Encyclopedia of Medical Devices and Instrumentation

Second Edition

John G. Webster

ENCYCLOPEDIA OF

MEDICAL DEVICES AND INSTRUMENTATION

Second Edition

VOLUME 1

Alloys, Shape Memory – Brachytherapy, Intravascular

ENCYCLOPEDIA OF MEDICAL DEVICES AND INSTRUMENTATION, SECOND EDITION

ENCYCLOPEDIA OF

MEDICAL DEVICES AND INSTRUMENTATION

Second Edition
Volume 1

Alloys, Shape Memory – Brachytherapy, Intravascular

Edited by

John G. Webster
University of Wisconsin–Madison

The *Encyclopedia of Medical Devices and Instrumentation* is available online at http://www.mrw.interscience.wiley.com/emdi

WILEY-INTERSCIENCE

A John Wiley & Sons, Inc., Publication

Published by John Wiley & Sons, Inc., Hoboken, New Jersey

Published simultaneously in Canada

Library of Congress Cataloging-in-Publication Data:

Encylopedia of medical devices & instrumentation/by John G. Webster,

editor in chief. – 2nd ed.
p. ; cm.
Rev. ed. of: Encyclopedia of medical devices and instrumentation. 1988.
Includes bibliographical references and index.
ISBN-13 978-0-471-26358-6 (set : cloth)
ISBN-10 0-471-26358-3 (set : cloth)
ISBN-13 978-0-470-04066-9 (v. 1 : cloth)
ISBN-10 0-470-04066-1 (v. 1 : cloth)

1. Medical instruments and apparatus–Encyclopedias. 2. Biomedical engineering–Encyclopedias. 3. Medical physics–Encyclopedias. 4. Medicine–Data processing–Encyclopedias. I. Webster, John G., 1932- . II. Title: Encyclopedia of medical devices and instrumentation.

[DNLM: 1. Equipment and Supplies–Encyclopedias–English. W 13

E555 2006]
R856.A3E53 2006
610.2803–dc22
2005028946

Printed in the United States of America

10 9 8 7 6 5 4 3 2 1

CONTRIBUTOR LIST

ABDEL HADY, MAZEN, *McMaster University, Hamilton, Ontario Canada*, Bladder Dysfunction, Neurostimulation of

ABEL, L.A., *University of Melbourne, Melbourne, Australia*, Ocular Motility Recording and Nystagmus

ABREU, BEATRIZ C., *Transitional Learning Center at Galveston, Galveston, Texas*, Rehabilitation, Computers in Cognitive

ALEXANDER, A.L., *University of Wisconsin–Madison, Madison, Wisconsin*, Magnetic Resonance Imaging

ALI, ABBAS, *University of Illinois, at Urbana-Champaign*, Bioinformatics

ALI, MÜFTÜ, *School of Dental Medicine, Boston, Massachusetts*, Tooth and Jaw, Biomechanics of

ALPERIN, NOAM, *University of Illinois at Chicago, Chicago, Illinois*, Hydrocephalus, Tools for Diagnosis and Treatment of

ANSON, DENIS, *College Misericordia, Dallas, Pennsylvania*, Environmental Control

ARENA, JOHN C., *VA Medical Center and Medical College of Georgia*, Biofeedback

ARIEL, GIDEON, *Ariel Dynamics, Canyon, California*, Biomechanics of Exercise Fitness

ARMSTRONG, STEVE, *University of Iowa, Iowa City, Iowa*, Biomaterials for Dentistry

ASPDEN, R.M., *University of Aberdeen, Aberdeen, United Kingdom*, Ligament and Tendon, Properties of

AUBIN, C.E., *Polytechniquie Montreal, Montreal Quebec, Canada*, Scoliosis, Biomechanics of

AYRES, VIRGINIA M., *Michigan State University, East Lansing, Michigan*, Microscopy, Scanning Tunneling

AZANGWE, G., Ligament and Tendon, Properties of

BACK, LLOYD H., *California Institute of Technology, Pasadena, California*, Coronary Angioplasty and Guidewire Diagnostics

BADYLAK, STEPHEN F., *McGowan Institute for Regenerative Medicine, Pittsburgh, Pennsylvania*, Sterilization of Biologic Scaffold Materials

BANDYOPADHYAY, AMIT, *Washington State University, Pullman, Washington*, Orthopedic Devices, Materials and Design for

BANERJEE, RUPAK K., *University of Cincinnati, Cincinnati, Ohio*, Coronary Angioplasty and Guidewire Diagnostics

BARBOUR, RANDALL L., *State University of New York Downstate Medical Center, Brooklyn, New York*, Peripheral Vascular Noninvasive Measurements

BARKER, STEVEN J., *University of Arizona, Tucson, Arizona*, Oxygen Monitoring

BARTH, ROLF F., *The Ohio State University, Columbus, Ohio*, Boron Neutron Capture Therapy

BECCHETTI, F.D., *University of Michigan, Ann Arbor, Michigan*, Radiotherapy, Heavy Ion

BELFORTE, GUIDO, *Politecnico di Torino – Department of Mechanics*, Laryngeal Prosthetic Devices

BENKESER, PAUL, *Georgia Institute of Technology, Atlanta, Georgia*, Biomedical Engineering Education

BENNETT, JAMES R., *University of Iowa, Iowa City, Iowa*, Digital Angiography

BERSANO-BEGEY, TOMMASO, *University of Michigan, Ann Arbor, Michigan*, Microbioreactors

BIGGS, PETER J., *Harvard Medical School, Boston, Massachusetts*, Radiotherapy, Intraoperative

BIYANI, ASHOK, *University of Toledo, and Medical College of Ohio, Toledo, Ohio*, Human Spine, Biomechanics of

BLOCK, W.F., *University of Wisconsin–Madison, Madison, Wisconsin*, Magnetic Resonance Imaging

BLUE, THOMAS E., *The Ohio State University, Columbus, Ohio*, Boron Neutron Capture Therapy

BLUMSACK, JUDITH T., *Disorders Auburn University, Auburn, Alabama*, Audiometry

BOGAN, RICHARD K., *University of South Carolina, Columbia, South Carolina*, Sleep Laboratory

BOKROS, JACK C., *Medical Carbon Research Institute, Austin, Texas*, Biomaterials, Carbon

BONGIOANNINI, GUIDO, *ENT Division Mauriziano Hospital, Torino, Italy*, Laryngeal Prosthetic Devices

BORAH, JOSHUA, *Applied Science Laboratories, Bedford, Massachusetts*, Eye Movement, Measurement Techniques for

BORDEN, MARK, *Director of Biomaterials Research, Irvine, California*, Biomaterials, Absorbable

BORTON, BETTIE B., *Auburn University Montgomery, Montgomery, Alabama*, Audiometry

BORTON, THOMAS E., *Auburn University Montgomery, Montgomery, Alabama*, Audiometry

BOSE SUSMITA,, *Washington State University, Pullman, Washington*, Orthopedic Devices, Materials and Design for

BOVA, FRANK J., *M. D. Anderson Cancer Center Orlando, Orlando, FL*, Radiosurgery, Stereotactic

BRENNER, DAVID J., *Columbia University Medical Center, New York, New York*, Computed Tomography Screening

BREWER, JOHN M., *University of Georgia*, Electrophoresis

BRIAN, L. DAVIS, *Lerner Research Institute, The Cleveland Clinic Foundation, Cleveland, Ohio*, Skin, Biomechanics of

BRITT, L.D., *Eastern Virginia Medical School, Norfolk, Virginia*, Gastrointestinal Hemorrhage

BRITT, R.C., *Eastern Virginia Medical School, Norfolk, Virginia*, Gastrointestinal Hemorrhage

BROZIK, SUSAN M., *Sandia National Laboratories, Albuquerque, New Mexico*, Microbial Detection Systems

BRUNER, JOSEPH P., *Vanderbilt University Medical Center, Nashville, Tennessee*, Intrauterine Surgical Techniques

BRUNSWIG NEWRING, KIRK A., *University of Nevada, Reno, Nevada*, Sexual Instrumentatio n

BRUYANT, PHILIPPE P., *University of Massachusetts, North Worcester, Massachusetts*, Nuclear Medicine, Computers in

BUNNELL, BERT J., *Bunnell Inc., Salt Lake City, Utah*, High Frequency Ventilation

CALKINS, JERRY M., *Defense Research Technologies, Inc., Rockville, Maryland*, Medical Gas Analyzers

CANNON, MARK, *Northwestern University, Chicago, Illinois*, Resin-Based Composites

CAPPELLERI, JOSEPH C., *Pfizer Inc., Groton, Connecticut*, Quality-of-Life Measures, Clinical Significance of

CARDOSO, JORGE, *University of Madeira, Funchal, Portugal*, Office Automation Systems

CARELLO, MASSIMILIANA, *Politecnicodi Torino – Department of Mechanics*, Laryngeal Prosthetic Devices

CASKEY, THOMAS C., *Cogene Biotech Ventures, Houston, Texas*, Polymerase Chain Reaction

CECCIO, STEVEN, *University of Michigan, Ann Arbor, Michigan*, Heart Valve Prostheses, In Vitro Flow Dynamics of

CHAN, JACKIE K., *Columbia University, New York, New York*, Photography, Medical

CHANDRAN, K.B., *University of Iowa, Iowa City, Iowa*, Heart Valve Prostheses

CHATZANDROULIS, S., *NTUA, Athens, Attiki, Greece*, Capacitive Microsensors for Biomedical Applications

CHAVEZ, ELIANA, *University of Pittsburgh, Pittsburgh, Pennsylvania*, Mobility Aids

CHEN, HENRY, *Stanford University, Palo Alto, California*, Exercise Stress Testing

CHEN, JIANDE, *University of Texas Medical Branch, Galveston, Texas*, Electrogastrogram

CHEN, YAN, *Lerner Research Institute, The Cleveland Clinic Foundation, Cleveland, Ohio*, Skin, Biomechanics of

CHEYNE, DOUGLAS, *Hospital for Sick Children Research Institute*, Biomagnetism

CHUI, CHEN-SHOU, *Memorial Sloan-Kettering Cancer Center, New York, New York*, Radiation Therapy Treatment Planning, Monte Carlo Calculations in

CLAXTON, NATHAN S., *The Florida State University, Tallahassee, Florida*, Microscopy, Confocal

CODERRE, JEFFREY A., *Massachus etts Institute of Technology, Cambridge, Massachusetts*, Boron Neutron Capture Therapy

COLLINS, BETH, *University of Mississippi Medical Center, Jackson, Mississippi*, Hyperbaric Medicine

COLLINS, DIANE, *University of Pittsburgh, Pittsburgh, Pennsylvania*, Mobility Aids

CONSTANTINOU, C., *Columbia University Radiation Oncology, New York, New York*, Phantom Materials in Radiology

COOK, ALBERT, *University of Alberta*, *Edmonton, Alberta, Canada*, Communication Devices

COOPER, RORY, *University of Pittsburgh, Pittsburgh, Pennsylvania*, Mobility Aids

CORK, RANDALL C., *Louisiana State University, Shreveport, Louisiana*, Monitoring, Umbilical Artery and Vein, Blood Gas Measurements; Transcuta neous Electrical Nerve Stimulation (TENS); Ambulatory Monitoring

COX, JOSEPHINE H., *Walter Reed Army Institute of Research, Rockville, Maryland*, Blood Collection and Processing

CRAIG, LEONARD, *Feinberg School of Medicine of Northwestern University, Chicago, Illinois*, Ventilators, Acute Medical Care

CRESS, CYNTHIA J., *University of Nebraska, Lincoln, Nebraska*, Communicative Disorders, Computer Applications for

CUMMING, DAVID R.S., *University of Glasgow, Glasgow, United Kingdom*, Ion-Sensitive Field-Effect Transistors

CUNNINGHAM, JOHN R., *Camrose, Alberta, Canada*, Cobalt 60 Units for Radiotherapy

D'ALESSANDRO, DAVID, *Montefiore Medical Center, Bronx, New York*, Heart–Lung Machines

D'AMBRA, MICHAEL N., *Harvard Medical School, Cambridge, Massachusetts*, Cardiac Output, Thermodilution Measurement of

DADSETAN, MAHROKH, *Mayo Clinic, College of Medicine, Rochester, Minnesota*, Microscopy, Electron

DALEY, MICHAEL L., *The University of Memphis, Memphis, Tennessee*, Monitoring, Intracranial Pressure

DAN, LOYD, *Linköping University, Linköping, Sweden*, Thermocouples

DAS, RUPAK, *University of Wisconsin, Madison, Wisconsin*, Brachytherapy, High Dosage Rate

DATTAWADKAR, AMRUTA M., *University of Wisconsin, Madison, Madison, Wisconsin*, Ocular Fundus Reflectometry

DAVIDSON, MICHAEL W., *The Florida State University, Tallahassee, Florida*, Microscopy, Confocal

DE LUCA, CARLO, *Boston University, Boston, Massachusetts*, Electromyography

DE SALLES, ANTONIO A.F., *UCLA Medical School, Los Angeles, California*, Stereotactic Surgery

DECAU, SABIN, *University of Maryland, School of Medicine*, Shock, Treatment of

DECHOW, PAUL C., *A & M University Health Science Center, Dallas, Texas*, Strain Gages

DELBEKE, JEAN, *Catholique University of Louvain, Brussels, Belgium*, Visual Prostheses

DELL'OSSO, LOUIS F., *Case Western Reserve University, Cleveland, Ohio*, Ocular Motility Recording and Nystagmus

DELORME, ARNAUD, *University of San Diego, La Jolla, California*, Statistical Methods

DEMENKOFF, JOHN, *Mayo Clinic, Scottsdale, Arizona*, Pulmonary Physiology

DEMIR, SEMAHAT S., *The University of Memphis and The University of Tennessee Health Science Center, Memphis, Tennessee*, Electrophysiology

DEMLING, ROBERT H., *Harvard Medical School*, Skin Substitute for Burns, Bioactive

DENNIS, MICHAEL J., *Medical University of Ohio, Toledo, Ohio*, Computed Tomography

DESANTI, LESLIE, *Harvard Medical School*, Skin Substitute for Burns, Bioactive

DEUTSCH, STEVEN, *Pennsylvania State University, University Park, Pennsylvania*, Flowmeters

DEVINENI, TRISHUL, *Conemaugh Health System*, Biofeedback

DI BELLA EDWARD, V.R., *University of Utah*, Tracer Kinetics

DIAKIDES, NICHOLAS A., *Advanced Concepts Analysis, Inc., Falls Church, Virginia*, Thermography

DOLAN, PATRICIA L., *Sandia National Laboratories, Albuquerque, New Mexico*, Microbial Detection Systems

DONOVAN, F.M., *University of South Alabama*, Cardiac Output, Indicator Dilution Measurement of

DOUGLAS, WILSON R., *Children's Hospital of Philadelphia, Philadelphia, Pennsylvania*, Intrauterine Surgical Techniques

DRAPER, CRISSA, *University of Nevada, Reno, Nevada*, Sexual Instrumentation

DRZEWIECKI, TADEUSZ M., *Defense Research Technologies, Inc., Rockville, Maryland*, Medical Gas Analyzers

DURFEE, W.K., *University of Minnesota, Minneapolis, Minnesota*, Rehabilitation and Muscle Testing

DYRO, JOSEPH F., *Setauket, New York*, Safety Program, Hospital

DYSON, MARY, *Herts, United Kingdom*, Heat and Cold, Therapeutic

ECKERLE, JOSEPH S., *SRI International, Menlo Park, California*, Tonometry, Arterial

EDWARDS, BENJAMIN, *University of Wisconsin-Madison, Madison, Wisconsin*, Codes and Regulations: Radiation

EDWARDS, THAYNE L., *University of Washington, Seattle, Washington*, Chromatography

EKLUND, ANDERS, *University of Illinois at Chicago, Chicago, Illinois*, Hydrocephalus, Tools for Diagnosis and Treatment of

EL SOLH, ALI A., *Erie County Medical Center, Buffalo, New York*, Sleep Studies, Computer Analysis of

ELMAYERGI, NADER, *McMaster University, Hamilton, Ontario, Canada*, Bladder Dysfunction, Neurostimulation of

ELSHARYDAH, AHMAD, *Louisiana State University, Baton Rouge, Louisiana*, Ambulatory Monitoring; Monitoring, Umbilical Artery and Vein, Blood Gas Measurements

FADDY, STEVEN C., *St. Vincents Hospital, Sydney, Darlinghurst, Australia*, Cardiac Output, Fick Technique for

FAHEY, FREDERIC H., *Childrens Hospital Boston*, Computed Tomography, Single Photon Emission

FAIN, S.B., *University of Wisconsin–Madison, Madison, Wisconsin*, Magnetic Resonance Imaging

FELDMAN, JEFFREY, *Childrens Hospital of Philadelphia, Philadelphia, Pennsylvania*, Anesthesia Machines

FELLERS, THOMAS J., *The Florida State University, Tallahassee, Florida*, Microscopy, Confocal

FERRARA, LISA, *Cleveland Clinic Foundation, Cleveland, Ohio*, Human Spine, Biomechanics of

FERRARI, MAURO, *The Ohio State University, Columbus, Ohio*, Drug Delivery Systems

FONTAINE, ARNOLD A., *Pennsylvania State University, University Park, Pennsylvania*, Flowmeters

FOUST, MILTON J., JR, *Medical University of South Carolina Psychiatry and Behavioral Sciences, Charleston, South Carolina*, Electroconvulsive Therapy

FRASCO, PETER, *Mayo Clinic Scottsdale, Scottsdale, Arizona*, Temperature Monitoring

FRAZIER, JAMES, *Louisiana State University, Baton Rouge, Louisiana*, Ambulatory Monitoring

FREIESLEBEN DE BLASIO, BIRGITTE, *University of Oslo, Oslo, Norway*, Impedance Spectroscopy

FRESTA, MASSIMO, *University of Catanzaro Magna Græcia, Germaneto (CZ), Italy*, Drug Delivery Systems

FREYTES, DONALD O., *McGowan Institute for Regenerative Medicine, Pittsburgh Pennsylvania*, Sterilization of Biologic Scaffold Materials

FROEHLICHER, VICTOR, *VA Medical Center, Palo Alto, California*, Exercise Stress Testing

FUNG, EDWARD K., *Columbia University, New York, New York*, Photography, Medical

GAGE, ANDREW A., *State University of New York at Buffalo, Buffalo, New York*, Cryosurgery

GAGLIO, PAUL J., *Columbia University College of Physicians and Surgeons*, Liver Transplantation

GARDNER, REED M., *LDS Hospital and Utah University, Salt Lake City, Utah*, Monitoring, Hemodynamic

GEJERMAN, GLEN, *Hackensack University Medical, Hackensack, New Jersey*, Radiation Therapy, Quality Assurance in

GEORGE, MARK S., *Medical University of South Carolina Psychiatry and Behavioral Sciences, Charleston, South Carolina*, Electroconvulsive Therapy

GHARIEB, R.R., *Infinite Biomedical Technologies, Baltimore, Maryland*, Neurological Monitors

GLASGOW, GLENN P., *Loyola University of Chicago, Maywood, Illinois*, Radiation Protection Instrumentation

GLASGOW, GLENN, *University of Wisconsin-Madison, Madison, Wisconsin*, Codes and Regulations: Radiation

GOEL, VIJAY K., *University of Toledo, and Medical College of Ohio, Toledo, Ohio*, Human Spine, Biomechanics of

GOETSCH, STEVEN J., *San Diego Gamma Knife Center, La Jolla, California*, Gamma Knife

GOLDBERG, JAY R., *Marquette University Milwaukee, Wisconsin*, Minimally Invasive Surgery

GOLDBERG, ZELENNA, *Department of Radiation Oncology, Davis, California*, Ionizing Radiation, Biological Effects of

GOPALAKRISHNAKONE, P., *National University of Singapore, Singapore*, Immunologically Sensitive Field-Effect Transistors

GOPAS, JACOB, *Ben Gurion University of the Negev, Beer Sheva, Israel*, Monoclonal Antibodies

GORGULHO, ALESSANDRA, *UCLA Medical School, Los Angeles, California*, Stereotactic Surgery

GOUGH, DAVID A., *University of California, La Jolla, California*, Glucose Sensors

GOUSTOURIDIS, D., *NTUA, Athens, Attiki, Greece*, Capacitive Microsensors for Biomedical Applications

GRABER, HARRY L., *State University of New York Downstate Medical Center, Brooklyn, New York*, Peripheral Vascular Noninvasive Measurements

GRAÇA, M., *Louisiana State University, Baton Rouge, Louisiana*, Boron Neutron Capture Therapy

GRANT, WALTER III, *Baylor College of Medicine, Houston, Texas*, Radiation Therapy, Intensity Modulated

GRAYDEN, EDWARD, *Mayo Health Center, Albertlea, Minnesota*, Cardiopulmonary Resuscitation

GREEN, JORDAN R., *University of Nebraska, Lincoln, Nebraska*, Communicative Disorders, Computer Applications for

HAEMMERICH, DIETER, *Medical University of South Carolina, Charleston, South Carolina*, Tissue Ablation

HAMAM, HABIB, *Université de Moncton, Moncton New Brunswick, Canada*, Lenses, Intraocular

HAMMOND, PAUL A., *University of Glasgow, Glasgow, United Kingdom*, Ion-Sensitive Field-Effect Transistors

HANLEY, JOSEPH, *Hackensack University Medical, Hackensack, New Jersey*, Radiation Therapy, Quality Assurance in

HARLEY, BRENDAN A., *Massachusetts Institute of Technology*, Skin Tissue Engineering for Regeneration

HARPER, JASON C., *Sandia National Laboratories, Albuquerque, New Mexico*, Microbial Detection Systems

HASMAN, ARIE, *Maastricht, The Netherlands*, Medical Education, Computers in

HASSOUNA, MAGDY, *Toronto Western Hospital, Toronto, Canada*, Bladder Dysfunction, Neurostimulation of

HAYASHI, KOZABURO, *Okayama University of Science, Okayama, Japan*, Arteries, Elastic Properties of

HENCH, LARRY L., *Imperial College London, London, United Kingdom*, Biomaterials: Bioceramics

HETTRICK, DOUGLAS A., *Sr. Principal Scientist Medtronic, Inc., Minneapolis, Minnesota*, Bioimpedance in Cardiovascular Medicine

HIRSCH-KUCHMA, MELISSA, *University of Central Florida NanoScience Technology Center, Orlando, Florida*, Biosurface Engineering

HOLDER, GRAHAM E., *Moorfields Eye Hospital, London, United Kingdom*, Electroretinography

HOLMES, TIMOTHY, *St. Agnes Cancer Center, Baltimore, Maryland*, Tomotherapy

HONEYMAN-BUCK, JANICE C., *University of Florida, Gainesville, Florida*, Radiology Information Systems

HOOPER, BRETT A., *Areté Associates, Arlington, Virginia*, Endoscopes

HORN, BRUCE, *Kaiser Permanente, Los Angeles, California*, X-Rays Production of

HORNER, PATRICIA I., *Biomedical Engineering Society Landover, Maryland*, Medical Engineering Societies and Organizations

HOROWITZ, PAUL M., *University of Texas, San Antonio, Texas*, Fluorescence Measurements

HOU, XIAOLIN, *Risø National Laboratory, Roskilde, Denmark*, Neutron Activation Analysis

HOVORKA, ROMAN, *University of Cambridge, Cambridge, United Kingdom*, Pancreas, Artificial

HUANG, H.K., *University of Southern California*, Teleradiology

HUNT, ALAN J., *University of Michigan, Ann Arbor, Michigan*, Optical Tweezers

HUTTEN, HELMUT, *University of Technology, Graz, Australia*, Impedance Plethysmography

IAIZZO, P.A., *University of Minnesota, Minneapolis, Minnesota*, Rehabilitation and Muscle Testing

IBBOTT, GEOFFREY S., *Anderson Cancer Center, Houston, Texas*, Radiation Dosimetry, Three-Dimensional

INGHAM, E., *University of Leeds, Leeds, United Kingdom*, Hip Joints, Artificial

ISIK, CAN, *Syracuse University, Syracuse, New York*, Blood Pressure Measurement

JAMES, SUSAN P., *Colorado State University, Fort Collins, Colorado*, Biomaterials: Polymers

JENSEN, WINNIE, *Aalborg University, Aalborg, Denmark*, Electroneurography

JIN, CHUNMING, *North Carolina State University, Raleigh, North Carolina*, Biomaterials, Corrosion and Wear of

JIN, Z.M., *University of Leeds, Leeds, United Kingdom*, Hip Joints, Artificial

JOHNSON, ARTHUR T., *University of Maryland College Park, Maryland*, Medical Engineering Societies and Organizations

JONES, JULIAN R., *Imperial College London, London, United Kingdom*, Biomaterials: Bioceramics

JOSHI, ABHIJEET, *Abbott Spine, Austin, Texas*, Spinal Implants

JUNG, RANU, *Arizona State University, Tempe, Arizona*, Functional Electrical Stimulation

JURISSON, SILVIA S., *University of Missouri Columbia, Missouri*, Radionuclide Production and Radioactive Decay

KAEDING, PATRICIA J., *Godfrey & Kahn S.C., Madison, Wisconsin*, Codes and Regulations: Medical Devices

KAMATH, CELIA C., *Mayo Clinic, Rochester, Minnesota*, Quality-of-Life Measures, Clinical Significance of

KANE, MOLLIE, *Madison, Wisconsin*, Contraceptive Devices

KATHERINE, ANDRIOLE P., *Harvard Medical School, Boston, Massachusetts*, Picture Archiving and Communication Systems

KATSAGGELOS, AGGELOS K., *Northwestern University, Evanston, Illinois*, DNA Sequencing

KATZ, J. LAWRENCE, *University of Missouri-Kansas City, Kansas City, Missouri*, Bone and Teeth, Properties of

KESAVAN, SUNIL, *Akebono Corporation, Farmington Hills, Michigan*, Linear Variable Differential Transformers

KHANG, GILSON, *Chonbuk National University*, Biomaterials: Tissue Engineering and Scaffolds

KHAODHIAR, LALITA, *Harvard Medical School, Boston, Massachusetts*, Cutaneous Blood Flow, Doppler Measurement of

KIM, MOON SUK, *Korea Research Institutes of Chemical Technology*, Biomaterials: Tissue Engineering and Scaffolds

KIM, YOUNG KON, *Inje University, Kimhae City, Korea*, Alloys, Shape Memory

KINDWALL, ERIC P., *St. Luke's Medical Center, Milwaukee, Wisconsin*, Hyperbaric Oxygenation

KING, MICHAEL A., *University of Massachusetts, North Worcester, Massachusetts*, Nuclear Medicine, Computers in

KLEBE, ROBERT J., *University of Texas, San Antonio, Texas*, Fluorescence Measurements

KLEIN, BURTON, *Burton Klein Associates, Newton, Massachusetts*, Gas and Vacuum Systems, Centrally Piped Medical

KNOPER, STEVEN R., *University of Arizona College of Medicine*, Ventilatory Monitoring

KONTAXAKIS, GEORGE, *Universidad Politécnica de Madrid, Madrid, Spain*, Positron Emission Tomography

KOTTKE-MARCHANT, KANDICE, *The Cleveland Clinic Foundation, Cleveland, Ohio*, Vascular Graft Prosthesis

KRIPFGANS, OLIVER, *University of Michigan, Ann Arbor, Michigan*, Ultrasonic Imaging

KULKARNI, AMOL D., *University of Wisconsin–Madison, Madison, Wisconsin*, Ocular Fundus Reflectometry, Visual Field Testing

KUMARADAS, J. CARL, *Ryerson University, Toronto, Ontario, Canada*, Hyperthermia, Interstitial

KUNICKA, JOLANTA, *Bayer HealthCare LLC, Tarrytown, New York*, Differential Counts, Automated

KWAK, KWANJ JOO, *University of Miami Miller School of Medicine, Miami, Florida*, Microscopy, Scanning Force

LAKES, RODERIC, *University of Wisconsin-Madison*, Bone and Teeth, Properties of

LAKKIREDDY, DHANUNJAYA, *The Cleveland Clinic Foundation, Cleveland, Ohio*, Hyperthermia, Ultrasonic

LARSEN, COBY, *Case Western Reserve University, Cleveland, Ohio*, Vascular Graft Prosthesis

LASTER, BRENDA H., *Ben Gurion University of the Negev, Beer Sheva, Israel*, Monoclonal Antibodies

LATTA, LOREN, *University of Miami, Coral Gables, Florida*, Rehabilitation, Orthotics in

LEDER, RON S., *Universidad Nacional Autonoma de Mexico Mexico, Distrito Federal*, Continuous Positive Airway Pressure

LEE, CHIN, *Harvard Medical School, Boston, Massachusetts*, Radiotherapy Treatment Planning, Optimization of; Hyperthermia, Interstitial

LEE, HAI BANG, *Korea Research Institutes of Chemical Technology*, Biomaterials: Tissue Engineering and Scaffolds

LEE, SANG JIN, *Korea Research Institutes of Chemical Technology*, Biomaterials: Tissue Engineering and Scaffolds

LEI, LIU, *Department of General Engineering, Urbana, Illinois*, Bioinformatics

LEI, XING, *Stanford University, Stanford, California*, Radiation Dose Planning, Computer-Aided

LEWIS, MATTHEW C., *Medical College of Wisconsin, Milwaukee, Wisconsin*, Hyperbaric Oxygenation

LI, CHAODI, *University of Notre Dame, Notre Dame, Indiana*, Bone Cement, Acrylic

LI, JONATHAN G., *University of Florida, Gainesville, Florida*, Radiation Dose Planning, Computer-Aided

LI, QIAO, *University of Michigan, Ann Arbor, Michigan*, Immunotherapy

LI, YANBIN, *University of Arkansas, Fayetteville, Arkansas*, Piezoelectric Sensors

LIBOFF, A.R., *Oakland University, Rochester, Michigan*, Bone Ununited Fracture and Spinal Fusion, Electrical Treatment of

LIGAS, JAMES, *University of Connecticut, Farmington, Connecticut*, Respiratory Mechanics and Gas Exchange

LIMOGE, AIME, *The René Descartes University of Paris, Paris, France*, Electroanalgesia, Systemic

LIN, PEI-JAN PAUL, *Beth Israel Deaconess Medical Center, Boston, Massachusets*, Mammography

LIN, ZHIYUE, *University of Kansas Medical Center, Kansas City, Kansas*, Electrogastrogram

LINEAWEAVER, WILLIAM C., *Unive rsity of Mississippi Medical Center, Jackson, Mississippi*, Hyperbaric Medicine

LIPPING, TARMO, *Tampere University of Technology, Pori, Finland*, Monitoring in Anesthesia

LIU, XIAOHUA, *The University of Michigan, Ann Arbor, Michigan*, Polymeric Materials

LLOYD, J.J., *Regional Medical Physics Department, Newcastle-upon-Tyne, United Kingdom*, Ultraviolet Radiation in Medicine

LOEB, ROBERT, *University of Arizona, Tuscon, Arizona*, Anesthesia Machines

LOPES DE MELO, PEDRO, *State University of Rio de Janeiro, Térreo Salas, Maracanã, Thermistors*

LOUDON, ROBERT G., Lung Sounds

LOW, DANIEL A., *Washington University School of Medicine, St. Louis, Missouri*, Radiation Therapy Simulator

LU, LICHUN, *Mayo Clinic, College of Medicine, Rochester, Minnesota*, Microscopy, Electron

LU, ZHENG FENG, *Columbia University, New York, New York*, Screen-Film Systems

LYON, ANDREW W., *University of Calgary, Calgary, Canada*, Flame Atomic Emission Spectrometry and Atomic Absorption Spectrometry

LYON, MARTHA E., *University of Calgary, Calgary, Canada*, Flame Atomic Emission Spectrometry and Atomic Absorption Spectrometry

MA, C-M CHARLIE, *Fox Chase Cancer Center, Philadelphia, Pennsylvania*, X-Ray Therapy Equipment, Low and Medium Energy

MACIA, NARCISO F., *Arizona State University at the Polytechnic Campus, Mesa, Arizona*, Pneumotachometers

MACKENZIE, COLIN F., *University of Maryland, School of Medicine*, Shock, Treatment of

MACKIE, THOMAS R., *University of Wisconsin, Madison, Wisconsin*, Tomotherapy

MADNANI, ANJU, *LSU Medical Centre, Shreveport, Louisiana*, Transcutaneous Electrical Nerve Stimulation (TENS)

MADNANI, SANJAY, *LSU Medical Centre, Shreveport, Louisiana*, Transcutaneous Electrical Nerve Stimulation (TENS)

MADSEN, MARK T., *University of Iowa, Iowa City, Iowa*, Anger Camera

MAGNANO, MAURO, *ENT Division Mauriziano Hospital, Torino, Italy*, Drug Delivery Systems

MANDEL, RICHARD, *Boston University School of Medicine, Boston, Massachusetts*, Colorimetry

MANNING, KEEFE B., *Pennsylvania State University, University Park, Pennsylvania*, Flowmeters

MAO, JEREMY J., *University of Illinois at Chicago, Chicago, Illinois*, Cartilage and Meniscus, Properties of

MARCOLONGO, MICHELE, *Drexel University, Philadelphia, Pennsylvania*, Spinal Implants

MAREK, MIROSLAV, *Georgia Institute of Technology, Atlanta, Georgia*, Biomaterials, Corrosion and Wear of

MARION, NICHOLAS W., *University of Illinois at Chicago, Chicago, Illinois*, Cartilage and Meniscus, Properties of

MASTERS, KRISTYN S., *University of Wisconsin, Madison, Wisconsin*, Tissue Engineering

MAUGHAN, RICHARD L., *Hospital of the University of Pennsylvania*, Neutron Beam Therapy

MCADAMS, ERIC, *University of Ulster at Jordanstown, Newtownabbey, Ireland*, Bioelectrodes

MCARTHUR, SALLY L., *University of Sheffield, Sheffield, United Kingdom*, Biomaterials, Surface Properties of

MCEWEN, MALCOM, *National Research Council of Canada, Ontario, Canada*, Radiation Dosimetry for Oncology

MCGOWAN, EDWARD J., *E.J. McGowan & Associates*, Biofeedback

MCGRATH, SUSAN, *Dartmouth College, Hanover, New Hampshire*, Oxygen Analyzers

MEEKS, SANFORD L., *University of Florida, Gainesville, Florida*, Radiosurgery, Stereotactic

MELISSA, PETER, *University of Central Florida NanoScience Technology Center, Orlando, Florida*, Biosurface Engineering

MENDELSON, YITZHAK, *Worcester Polytechnic Institute*, Optical Sensors

METZKER, MICHAEL L., *Baylor College of Medicine, Houston, Texas*, Polymerase Chain Reaction

MEYEREND, M.E., *University of Wisconsin–Madison, Madison, Wisconsin*, Magnetic Resonance Imaging

MICHLER, ROBERT, *Montefiore Medical Center, Bronx, New York*, Heart–Lung Machines

MICIC, MIODRAG, *MP Biomedicals LLC, Irvine, California*, Microscopy and Spectroscopy, Near-Field

MILLER, WILLIAM, *University of Missouri Columbia, Missouri*, Radionuclide Production and Radioactive Decay

MITTRA, ERIK, *Stony Brook University, New York*, Bone Density Measurement

MODELL, MARK, *Harvard Medical School, Boston, Massachusetts*, Fiber Optics in Medicine

MORE, ROBERT B., *RBMore Associates, Austin, Texas* Biomaterials Carbon

MORE, ROBERT, *Austin, Texas*, Heart Valves, Prosthetic

MORROW, DARREN, *Royal Adelaide Hospital, Adelaide, Australia*, Intraaortic Balloon Pump

MOURTADA, FIRAS, *MD Anderson Cancer Center, Houston, Texas*, Brachytherapy, Intravascular

MOY, VINCENT T., *University of Miami, Miller School of Medicine, Miami, Florida*, Microscopy, Scanning Force

MÜFTÜ, SINAN, *Northeastern University, Boston, Massachusetts*, Tooth and Jaw, Biomechanics of

MURPHY, RAYMOND L.H., Lung Sounds

MURPHY, WILLIAM L., *University of Wisconsin, Madison, Wisconsin*, Tissue Engineering

MURRAY, ALAN, *Newcastle University Medical Physics, Newcastle upon Tyne, United Kingdom*, Pace makers

MUTIC, SASA, *Washington University School of Medicine, St. Louis, Missouri*, Radiation Therapy Simulator

NARAYAN, ROGER J., *University of North Carolina, Chapel Hill, North Carolina*, Biomaterials, Corrosion and Wear of

NATALE, ANDREA, *The Cleveland Clinic Foundation, Cleveland, Ohio*, Hyperthermia, Ultrasonic

NAZERAN, HOMER, *The University of Texas, El Paso, Texas*, Electrocardiography, Computers in

NEUMAN, MICHAEL R., *Michigan Technological University, Houghton, Houghton, Michigan*, Fetal Monitoring, Neonatal Monitoring

NEUZIL, PAVEL, *Institute of Bioengineering and Nanotechnology, Singapore*, Immunologically Sensitive Field-Effect Transistors

NICKOLOFF, EDWARD L., *Columbia University, New York, New York*, X-Ray Quality Control Program

NIEZGODA, JEFFREY A., *Medical College of Wisconsin, Milwaukee, Wisconsin*, Hyperbaric Oxygenation

NISHIKAWA, ROBERT M., *The University of Chicago, Chicago, Illinois*, Computer-Assisted Detection and Diagnosis

NUTTER, BRIAN, *Texas Tech University, Lubbock, Texas*, Medical Records, Computers in

O'DONOHUE, WILLIAM, *University of Nevada, Reno, Nevada*, Sexual Instrumentation

ORTON, COLIN, *Harper Hospital and Wayne State University, Detroit, Michigan*, Medical Physics Literature

OZCELIK, SELAHATTIN, *Texas A&M University, Kingsville, Texas*, Drug Infusion Systems

PANITCH, ALYSSA, *Arizona State University, Tempe, Arizona*, Biomaterials: An Overview

PAOLINO, DONATELLA, *University of Catanzaro Magna Græcia, Germaneto (CZ), Italy*, Drug Delivery Systems

PAPAIOANNOU, GEORGE, *University of Wisconsin, Milwaukee, Wisconsin*, Joints, Biomechanics of

PARK, GRACE E., *Purdue University, West Lafayette, Indiana*, Porous Materials for Biological Applications

PARMENTER, BRETT A., *State University of New York at Buffalo, Buffalo, New York*, Sleep Studies, Computer Analysis of

PATEL, DIMPI, *The Cleveland Clinic Foundation, Cleveland, Ohio*, Hyperthermia, Ultrasonic

PEARCE, JOHN, *The University of Texas, Austin, Texas*, Electrosurgical Unit (ESU)

PELET, SERGE, *Massachusetts Institute of Technology, Cambridge, Massachusetts*, Microscopy, Fluorescence

PERIASAMY, AMMASI, *University of Virginia, Charlottesville, Virginia*, Cellular Imaging

PERSONS, BARBARA L., *University of Mississippi Medical Center, Jackson, Mississippi*, Hyperbaric Medicine

PIPER, IAN, *The University of Memphis, Memphis, Tennessee*, Monitoring, Intracranial Pressure

POLETTO, CHRISTOPHER J., *National Institutes of Health*, Tactile Stimulation

PREMINGER, GLENN M., *Duke University Medical Center, Durham, North Carolina*, Lithotripsy

PRENDERGAST, PATRICK J., *Trinity College, Dublin, Ireland*, Orthopedics, Prosthesis Fixation for

PREVITE, MICHAEL, *Massachusetts Institute of Technology, Cambridge, Massachusetts*, Microscopy, Fluorescence

PURDY, JAMES A., *UC Davis Medical Center, Sacramento, California*, Radiotherapy Accessories

QI, HAIRONG, *Advanced Concepts Analysis, Inc., Falls Church, Virginia*, Thermography

QIN, YIXIAN, *Stony Brook University, New York*, Bone Density Measurement

QUAN, STUART F., *University of Arizona, Tucson, Arizona*, Ventilatory Monitoring

QUIROGA, RODRIGO QUIAN, *University of Leicester, Leicester, United Kingdom*, Evoked Potentials

RAHAGHI, FARBOD N., *University of California, La Jolla, California*, Glucose Sensors

RAHKO, PETER S., *University of Wisconsin Medical School*, Echocardiography and Doppler Echocardiography

RALPH, LIETO, *University of Wisconsin–Madison, Madison, Wisconsin*, Codes and Regulations: Radiation

RAMANATHAN, LAKSHMI, *Mount Sinai Medical Center*, Analytical Methods, Automated

RAO, SATISH S.C., *University of Iowa College of Medicine, Iowa City, Iowa*, Anorectal Manometry

RAPOPORT, DAVID M., *NYU School of Medicine, New York, New York*, Continuous Positive Airway Pressure

REBELLO, KEITH J., *The Johns Hopkins University Applied Physics Lab, Laurel, Maryland*, Micro surgery

REDDY, NARENDER, *The University of Akron, Akron, Ohio*, Linear Variable Differential Transformers

REN-DIH, SHEU, *Memorial Sloan-Kettering Cancer Center, New York, New York*, Radiation Therapy Treatment Planning, Monte Carlo Calculations in

RENGACHARY, SETTI S., *Detroit, Michigan*, Human Spine, Biomechanics of

REPPERGER, DANIEL W., *Wright-Patterson Air Force Base, Dayton, Ohio*, Human Factors in Medical Devices

RITCHEY, ERIC R., *The Ohio State University, Columbus, Ohio*, Contact Lenses

RIVARD, MARK J., *Tufts New England Medical Center, Boston, Massachusetts*, Imaging Devices

ROBERTSON, J. DAVID, *University of Missouri, Columbia, Missouri*, Radionuclide Production and Radioactive Decay

ROTH, BRADLEY J., *Oakland University, Rochester, Michigan*, Defibrillators

ROWE-HORWEGE, R. WANDA, *University of Texas Medical School, Houston, Texas*, Hyperthermia, Systemic

RUMSEY, JOHN W., *University of Central Florida, Orlando, Florida*, Biosurface Engineering

RUTKOWSKI, GREGORY E., *University of Minnesota, Duluth, Minnesota*, Engineered Tissue

SALATA, O.V., *University of Oxford, Oxford, United Kingdom*, Nanoparticles

SAMARAS, THEODOROS, *Aristotle University of Thessaloniki Department of Physics, Thessaloniki, Greece*, Thermometry

SANGOLE, ARCHANA P., *Transitional Learning Center at Galveston, Galveston, Texas*, Rehabilitation, Computers in Cognitive

SARKOZI, LASZLO, *Mount Sinai School of Medicine*, Analytical Methods, Automated

SCHEK, HENRY III, *University of Michigan, Ann Arbor, Michigan*, Optical Tweezers

SCHMITZ, CHRISTOPH H., *State University of New York Downstate Medical Center, Brooklyn, New York*, Peripheral Vascular Noninvasive Measurements

SCHUCKERS, STEPHANIE A.C., *Clarkson University, Potsdam, New York*, Arrhythmia Analysis, Automated

SCOPE, KENNETH, *Northwestern University, Chicago, Illinois*, Ventilators, Acute Medical Care

SCOTT, ADZICK N., *University of Pennsylvania, Philadelphia, Pennsylvania*, Intrauterine Surgical Techniques

SEAL, BRANDON L., *Arizona State University, Tempe, Arizona*, Biomaterials: An Overview

SEALE, GARY, *Transitional Learning Center at Galveston, Galveston, Texas*, Rehabilitation, Computers in Cognitive

SEGERS, PATRICK, *Ghent University, Belgium*, Hemodynamics

SELIM, MOSTAFA A., *Cleveland Metropolitan General Hospital, Palm Coast, Florida*, Colposcopy

SETHI, ANIL, *Loyola University Medical Center, Maywood, Illinois*, X-Rays: Interaction with Matter

SEVERINGHAUS, JOHN W., *University of California in San Francisco*, CO_2 Electrodes

SHALODI, ABDELWAHAB D., *Cleveland Metropolitan General Hospital, Palm Coast, Florida*, Colposcopy

SHANMUGASUNDARAM, SHOBANA, *New Jersey Institute of Technology, Newark, New Jersey*, Polymeric Materials

SHARD, ALEXANDER G., *University of Sheffield, Sheffield United Kingdom*, Biomaterials, Surface Properties of

SHEN, LI-JIUAN, *National Taiwan University School of Pharmacy, Taipei, Taiwan*, Colorimetry

SHEN, WEI-CHIANG, *University of Southern California School of Pharmacy, Los Angeles, California*, Colorimetry

SHERAR, MICHAEL D., *London Health Sciences Centre and University of Western Ontario, London, Ontario, Canada*, Hyperthermia, Interstitial

SHERMAN, DAVID, *The Johns Hopkins University, Baltimore, Maryland*, Electroencephalography

SHI, DONGLU, *University of Cincinnati, Cincinnati, Ohio*, Biomaterials, Testing and Structural Properties of

SHUCARD, DAVID W.M., *State University of New York at Buffalo, Buffalo, New York*, Sleep Studies, Computer Analysis of

SIEDBAND, MELVIN P., *University of Wisconsin, Madison, Wisconsin*, Image Intensifiers and Fluoroscopy

SILBERMAN, HOWARD, *University of Southern California, Los Angeles, California*, Nutrition, Parenteral

SILVERMAN, GORDON, *Manhattan College*, Computers in the Biomedical Laboratory

SILVERN, DAVID A., *Medical Physics Unit, Rabin Medical Center, Petah Tikva, Israel*, Prostate Seed Implants

SINHA, PIYUSH, *The Ohio State University, Columbus, Ohio*, Drug Delivery Systems

SINHA, ABHIJIT ROY, *University of Cincinnati, Cincinnati, Ohio*, Coronary Angioplasty and Guidewire Diagnostics

SINKJÆR, THOMAS, *Aalborg University, Aalborg, Denmark*, Electroneurography

SLOAN, JEFFREY A., *Mayo Clinic, Rochester, Minnesota*, Quality-of-Life Measures, Clinical Significance of

SO, PETER T.C., *Massachusetts Institute of Technology, Cambridge, Massachusetts*, Microscopy, Fluorescence

SOBOL, WLAD T., *University of Alabama at Birmingham Health System, Birmingham, Alabama*, Nuclear Magnetic Resonance Spectroscopy

SOOD, SANDEEP, *University of Illinois at Chicago, Chicago, Illinois*, Hydrocephalus, Tools for Diagnosis and Treatment of

SPECTOR, MYRON, *Brigham and Women's Hospital, Boston, Massachusetts*, Biocompatibility of Materials

SPELMAN, FRANCIS A., *University of Washington*, Cochlear Prostheses

SRINIVASAN, YESHWANTH, *Texas Tech University, Lubbock, Texas*, Medical Records, Computers in

SRIRAM, NEELAMEGHAM, *University of Buffalo, Buffalo, New York*, Cell Counters, Blood

STARKO, KENTON R., *Point Roberts, Washington*, Physiological Systems Modeling

STARKSCHALL, GEORGE, *The University of Texas*, Radiotherapy, Three-Dimensional Conformal

STAVREV, PAVEL, *Cross Cancer Institute, Edmonton, Alberta, Canada*, Radiotherapy Treatment Planning, Optimization of

STENKEN, JULIE A., *Rensselaer Polytechnic Institute, Troy, New York*, Microdialysis Sampling

STIEFEL, ROBERT, *University of Maryland Medical Center, Baltimore, Maryland*, Equipment Acquisition

STOKES, I.A.F., *Polytechniquie Montreal, Montreal Quebec, Canada*, Scoliosis, Biomechanics of

STONE, M.H., *University of Leeds, Leeds, United Kingdom*, Hip Joints, Artificial

SU, XIAO-LI, *BioDetection Instruments LLC, Fayetteville, Arkansas*, Piezoelectric Sensors

SUBHAN, ARIF, *Masterplan Technology Management, Chatsworth, California*, Equipment Maintenance, Biomedical

SWEENEY, JAMES D., *Arizona State University, Tempe, Arizona*, Functional Electrical Stimulation

SZETO, ANDREW Y.J., *San Diego State University, San Diego, California*, Blind and Visually Impaired, Assistive Technology for

TAKAYAMA, SHUICHI, *University of Michigan, Ann Arbor, Michigan*, Microbioreactors

TAMUL, PAUL C., *Northwestern University, Chicago, Illinois*, Ventilators, Acute Medical Care

TAMURA, TOSHIYO, *Chiba University School of Engineering, Chiba, Japan*, Home Health Care Devices

TANG, XIANGYANG, *GE Healthcare Technologies, Wankesha, Wisconsin*, Computed Tomography Simulators

TAYLOR, B.C., *The University of Akron, Akron, Ohio*, Cardiac Output, Indicator Dilution Measurement of

TEMPLE, RICHARD O., *Transitional Learning Center at Galveston, Galveston, Texas*, Rehabilitation, Computers in Cognitive

TEN, STANLEY, *Salt Lake City, Utah*, Electroanalgesia, Systemic

TERRY, TERESA M., *Walter Reed Army Institute of Research, Rockville, Maryland*, Blood Collection and Processing

THAKOR, N.V., *Johns Hopkins University, Baltimore, Maryland*, Neurological Monitors

THIERENS, HUBERT M.A., *University of Ghent, Ghent, Belgium*, Radiopharmaceutical Dosimetry

THOMADSEN, BRUCE, *University of Wisconsin–Madison, Madison, Wisconsin*, Codes and Regulations: Radiation

TIPPER, J.L., *University of Leeds, Leeds, United Kingdom*, Hip Joints, Artificial

TOGAWA, TATSUO, *Waseda University, Saitama, Japan*, Integrated Circuit Temperature Sensor

TORNAI, MARTIN, *Duke University, Durham, North Carolina*, X-Ray Equipment Design

TRAN-SON-TAY, ROGER, *University of Florida, Gainesville, Florida*, Blood Rheology

YENI, YENER N., *Henry Ford Hospital, Detroit, Michigan*, Joints, Biomechanics of

YLI-HANKALA, ARVI, *Tampere University of Technology, Pori, Finland*, Monitoring in Anesthesia

YOKO, KAMOTANI, *University of Michigan, Ann Arbor, Michigan*, Microbioreactors

YOON, KANG JI, *Korea Institute of Science and Technology, Seoul, Korea*, Micropower for Medical Applications

YORKE, ELLEN, *Memorial Sloan-Kettering Cancer Center, New York, New York*, Radiation Therapy Treatment Planning, Monte Carlo Calculations in

YOSHIDA, KEN, *Aalborg University, Aalborg, Denmark*, Electroneurography

YOUNGSTEDT, SHAWN D., *University of South Carolina, Columbia, South Carolina*, Sleep Laboratory

YU, YIH-CHOUNG, *Lafayette College, Easton, Pennsylvania*, Blood Pressure, Automatic Control of

ZACHARIAH, EMMANUEL S., *University of Medicine and Dentistry of New Jersey, New Brunswick, New Jersey*, Immunologically Sensitive Field-Effect Transistors

ZAIDER, MARCO, *Memorial Sloan Kettering Cancer Center, New York, New York*, Prostate Seed Implants

ZAPANTA, CONRAD M., *Penn State College of Medicine, Hershey, Pennsylvania*, Heart, Artificial

ZARDENETA, GUSTAVO, *University of Texas, San Antonio, Texas*, Fluorescence Measurements

ZELMANOVIC, DAVID, *Bayer HealthCare LLC, Tarrytown, New York*, Differential Counts, Automated

ZHANG, MIN, *University of Washington, Seattle, Washington*, Biomaterials: Polymers

ZHANG, YI, *University of Buffalo, Buffalo, New York*, Cell Counters, Blood

ZHU, XIAOYUE, *University of Michigan, Ann Arbor, Michigan*, Microbioreactors

ZIAIE, BABAK, *Purdue University, W. Lafayette, Indiana*, Biotelemetry

ZIELINSKI, TODD M., *Medtronic, Inc., Minneapolis, Minnesota*, Bioimpedance in Cardiovascular Medicine

ZIESSMAN, HARVEY A., *Johns Hopkins University*, Computed Tomography, Single Photon Emission

PREFACE

This six-volume work is an alphabetically organized compilation of almost 300 articles that describe critical aspects of medical devices and instrumentation.

It is comprehensive. The articles emphasize the contributions of engineering, physics, and computers to each of the general areas of anesthesiology, biomaterials, burns, cardiology, clinical chemistry, clinical engineering, communicative disorders, computers in medicine, critical care medicine, dermatology, dentistry, ear, nose, and throat, emergency medicine, endocrinology, gastroenterology, genetics, geriatrics, gynecology, hematology, heptology, internal medicine, medical physics, microbiology, nephrology, neurology, nutrition, obstetrics, oncology, ophthalmology, orthopedics, pain, pediatrics, peripheral vascular disease, pharmacology, physical therapy, psychiatry, pulmonary medicine, radiology, rehabilitation, surgery, tissue engineering, transducers, and urology.

The discipline is defined through the synthesis of the core knowledge from all the fields encompassed by the application of engineering, physics, and computers to problems in medicine. The articles focus not only on what is now useful but also on what is likely to be useful in future medical applications.

These volumes answer the question, "What are the branches of medicine and how does technology assist each of them?" rather than "What are the branches of technology and how could each be used in medicine?" To keep this work to a manageable length, the practice of medicine that is unassisted by devices, such as the use of drugs to treat disease, has been excluded.

The articles are accessible to the user; each benefits from brevity of condensation instead of what could easily have been a book-length work. The articles are designed not for peers, but rather for workers from related fields who wish to take a first look at what is important in the subject.

The articles are readable. They do not presume a detailed background in the subject, but are designed for any person with a scientific background and an interest in technology. Rather than attempting to teach the basics of physiology or Ohm's law, the articles build on such basic concepts to show how the worlds of life science and physical science meld to produce improved systems. While the ideal reader might be a person with a Master's degree in biomedical engineering or medical physics or an M.D. with a physical science undergraduate degree, much of the material will be of value to others with an interest in this growing field. High school students and hospital patients can skip over more technical areas and still gain much from the descriptive presentations.

The *Encyclopedia of Medical Devices and Instrumentation* is excellent for browsing and searching for those new divergent associations that may advance work in a peripheral field. While it can be used as a reference for facts, the articles are long enough that they can serve as an educational instrument and provide genuine understanding of a subject.

One can use this work just as one would use a dictionary, since the articles are arranged alphabetically by topic. Cross references assist the reader looking for subjects listed under slightly different names. The index at the end leads the reader to all articles containing pertinent information on any subject. Listed on pages xxi to xxx are all the abbreviations and acronyms used in the *Encyclopedia.* Because of the increasing use of SI units in all branches of science, these units are provided throughout the *Encyclopedia* articles as well as on pages xxxi to xxxv in the section on conversion factors and unit symbols.

I owe a great debt to the many people who have contributed to the creation of this work. At John Wiley & Sons, Encyclopedia Editor George Telecki provided the idea and guiding influence to launch the project. Sean Pidgeon was Editorial Director of the project. Assistant Editors Roseann Zappia, Sarah Harrington, and Surlan Murrell handled the myriad details of communication between publisher, editor, authors, and reviewers and stimulated authors and reviewers to meet necessary deadlines.

My own background has been in the electrical aspects of biomedical engineering. I was delighted to have the assistance of the editorial board to develop a comprehensive encyclopedia. David J. Beebe suggested cellular topics such as microfluidics. Jerry M. Calkins assisted in defining the chemically related subjects, such as anesthesiology. Michael R. Neuman suggested subjects related to sensors, such as in his own work—neonatology. Joon B. Park has written extensively on biomaterials and suggested related subjects. Edward S. Sternick provided many suggestions from medical physics. The Editorial Board was instrumental both in defining the list of subjects and in suggesting authors.

This second edition brings the field up to date. It is available on the web at http://www.mrw.interscience.wiley.com/emdi, where articles can be searched simultaneously to provide rapid and comprehensive information on all aspects of medical devices and instrumentation.

JOHN G. WEBSTER
University of Wisconsin, Madison

LIST OF ARTICLES

ABBREVIATIONS AND ACRONYMS

AAMI	Association for the Advancement of Medical Instrumentation
AAPM	American Association of Physicists in Medicine
ABC	Automatic brightness control
ABET	Accreditation board for engineering training
ABG	Arterial blood gases
ABLB	Alternative binaural loudness balance
ABS	Acrylonitrile–butadiene–styrene
ac	Alternating current
AC	Abdominal circumference; Affinity chromatography
ACA	Automated clinical analyzer
ACES	Augmentative communication evaluation system
ACL	Anterior chamber lens
ACLS	Advanced cardiac life support
ACOG	American College of Obstetrics and Gynecology
ACR	American College of Radiology
ACS	American Cancer Society; American College of Surgeons
A/D	Analog-to-digital
ADC	Agar diffusion chambers; Analog-to-digital converter
ADCC	Antibody-dependent cellular cytotoxicity
ADCL	Accredited Dosimetry Calibration Laboratories
ADP	Adenosine diphosphate
A-D-T	Admission, discharge, and transfer
AE	Anion exchange; Auxiliary electrode
AEA	Articulation error analysis
AEB	Activation energy barrier
AEC	Automatic exposure control
AED	Automatic external defibrillator
AEMB	Alliance for Engineering in Medicine and Biology
AES	Auger electron spectroscopy
AESC	American Engineering Standards Committee
AET	Automatic exposure termination
AFO	Ankle-foot orthosis
AGC	Automatic gain control
AHA	American Heart Association
AI	Arterial insufficiency
AICD	Automatic implantable cardiac defibrillator
AID	Agency for International Development
AIDS	Acquired immune deficiency syndrome
AL	Anterior leaflet
ALG	Antilymphocyte globulin
ALS	Advanced life support; Amyotropic lateral sclerosis
ALT	Alanine aminotransferase
ALU	Arithmetic and logic unit
AM	Amplitude modulation
AMA	American Medical Association
amu	Atomic mass units
ANOVA	Analysis of variance
ANSI	American National Standards Institute
AP	Action potential; Alternative pathway; Anteroposterior
APD	Anterioposterior diameter
APL	Adjustable pressure limiting valve; Applied Physics Laboratory
APR	Anatomically programmed radiography
AR	Amplitude reduction; Aortic regurgitation; Autoregressive
Ara-C	Arabinosylcytosine
ARD	Absorption rate density
ARDS	Adult respiratory distress syndrome
ARGUS	Arrhythmia guard system
ARMA	Autoregressive-moving-average model
ARMAX	Autoregressive-moving-average model with external inputs
AS	Aortic stenosis
ASA	American Standards Association
ASCII	American standard code for information interchange
ASD	Antisiphon device
ASHE	American Society for Hospital Engineering
ASTM	American Society for Testing and Materials
AT	Adenosine-thiamide; Anaerobic threshold; Antithrombin
ATA	Atmosphere absolute
ATLS	Advanced trauma life support
ATN	Acute tubular necrosis
ATP	Adenosine triphosphate
ATPD	Ambient temperature pressure dry
ATPS	Ambient temperature pressure saturated
ATR	Attenuated total reflection
AUC	Area under curve
AUMC	Area under moment curve
AV	Atrioventricular
AZT	Azido thymidine
BA	Biliary atresia
BAEP	Brainstem auditory evoked potential
BAPN	Beta-amino-proprionitryl
BAS	Boston anesthesis system
BASO	Basophil
BB	Buffer base
BBT	Basal body temperature

BCC	Body-centered cubic
BCD	Binary-coded decimal
BCG	Ballistocardiogram
BCLS	Basic cardiac life support
BCRU	British Commitee on Radiation Units and Measurements
BDI	Beck depression inventory
BE	Base excess; Binding energy
BET	Brunauer, Emmett, and Teller methods
BH	His bundle
BI	Biological indicators
BIH	Beth Israel Hospital
BIPM	International Bureau of Weights and Measurements
BJT	Bipolar junction transistor
BMDP	Biomedical Programs
BME	Biomedical engineering
BMET	Biomedical equipment technician
BMO	Biomechanically optimized
BMR	Basal metabolic rate
BOL	Beginning of life
BP	Bereitschafts potential; Break point
BR	Polybutadiene
BRM	Biological response modifier
BRS	Bibliographic retrieval services
BSS	Balanced salt solution
BTG	Beta thromboglobulin
BTPS	Body temperature pressure saturated
BUN	Blood urea nitrogen
BW	Body weight
CA	Conductive adhesives
CABG	Coronary artery by-pass grafting
CAD/CAM	Computer-aided design/computer-aided manufacturing
CAD/D	Computer-aided drafting and design
CADD	Central axis depth dose
CAI	Computer assisted instruction; Computer-aided instruction
CAM	Computer-assisted management
cAMP	Cyclic AMP
CAPD	Continuous ambulatory peritoneal dialysis
CAPP	Child amputee prosthetic project
CAT	Computerized axial tomography
CATS	Computer-assisted teaching system; Computerized aphasia treatment system
CAVH	Continuous arteriovenous hemofiltration
CB	Conjugated bilirubin; Coulomb barrier
CBC	Complete blood count
CBF	Cerebral blood flow
CBM	Computer-based management
CBV	Cerebral blood volume
CC	Closing capacity
CCC	Computer Curriculum Company
CCD	Charge-coupled device
CCE	Capacitance contact electrode
CCF	Cross-correlation function
CCL	Cardiac catheterization laboratory
CCM	Critical care medical services
CCPD	Continuous cycling peritoneal dialysis
CCTV	Closed circuit television system
CCU	Coronary care unit; Critical care unit
CD	Current density
CDR	Complimentary determining region
CDRH	Center for Devices and Radiological Health
CEA	Carcinoembryonic antigen
CF	Conversion factor; Cystic fibrosis
CFC	Continuous flow cytometer
CFR	Code of Federal Regulations
CFU	Colony forming units
CGA	Compressed Gas Association
CGPM	General Conference on Weights and Measures
CHO	Carbohydrate
CHO	Chinese hamster ovary
CI	Combination index
CICU	Cardiac intensive care unit
CIF	Contrast improvement factor
CIN	Cervical intraepithelial neoplasia
CK	Creatine kinase
CLAV	Clavicle
CLSA	Computerized language sample analysis
CM	Cardiomyopathy; Code modulation
CMAD	Computer managed articulation diagnosis
CMI	Computer-managed instruction
CMRR	Common mode rejection ratio
CMV	Conventional mechanical ventilation; Cytomegalovirus
CNS	Central nervous system
CNV	Contingent negative variation
CO	Carbon monoxide; Cardiac output
COBAS	Comprehensive Bio-Analysis System
COPD	Chronic obstructive pulmonary disease
COR	Center of rotation
CP	Cerebral palsy; Closing pressure; Creatine phosphate
CPB	Cardiopulmonary bypass
CPET	Cardiac pacemaker electrode tips
CPM	Computerized probe measurements
CPP	Cerebral perfusion pressure; Cryoprecipitated plasma
CPR	Cardiopulmonary resuscitation
cps	Cycles per second
CPU	Central Processing unit
CR	Center of resistance; Conditioned response; Conductive rubber; Creatinine
CRBB	Complete right bundle branch block
CRD	Completely randomized design
CRL	Crown rump length
CRT	Cathode ray tube
CS	Conditioned stimulus; Contrast scale; Crown seat
CSA	Compressed spectral array
CSF	Cerebrospinal fluid
CSI	Chemical shift imaging
CSM	Chemically sensitive membrane
CT	Computed tomography; Computerized tomography
CTI	Cumulative toxicity response index
CV	Closing volume

C.V.	Coefficient of variation
CVA	Cerebral vascular accident
CVP	Central venous pressure
CVR	Cardiovascular resistance
CW	Continuous wave
CWE	Coated wire electrodes
CWRU	Case Western Reserve University
DAC	Digital-to-analog converter
DAS	Data acquisition system
dB	Decibel
DB	Direct body
DBMS	Data base management system
DBS	Deep brain stimulation
dc	Direct current
DCCT	Diabetes control and complications trial
DCP	Distal cavity pressure
DCS	Dorsal column stimulation
DDC	Deck decompression chamber
DDS	Deep diving system
DE	Dispersive electrode
DEN	Device experience network
DERS	Drug exception ordering system
DES	Diffuse esophageal spasm
d.f.	Distribution function
DHCP	Distributed Hospital Computer Program
DHE	Dihematoporphyrin ether
DHEW	Department of Health Education and Welfare
DHHS	Department of Health and Human Services
DHT	Duration of hypothermia
DI	Deionized water
DIC	Displacement current
DIS	Diagnostic interview schedule
DL	Double layer
DLI	Difference lumen for intensity
DM	Delta modulation
DME	Dropping mercury electrode
DN	Donation number
DNA	Deoxyribonucleic acid
DOF	Degree of freedom
DOS	Drug ordering system
DOT-NHTSA	Department of Transportation Highway Traffic Safety Administration
DPB	Differential pencil beam
DPG	Diphosphoglycerate
DQE	Detection quantum efficiency
DRESS	Depth-resolved surface coil spectroscopy
DRG	Diagnosis-related group
DSA	Digital subtraction angiography
DSAR	Differential scatter-air ratio
DSB	Double strand breaks
DSC	Differential scanning calorimetry
D-T	Deuterium-on-tritium
DTA	Differential thermal analysis
d.u.	Density unit
DUR	Duration
DVT	Deep venous thrombosis
EA	Esophageal accelerogram
EB	Electron beam
EBCDIC	Extended binary code decimal interchange code
EBS	Early burn scar
EBV	Epstein–Barr Virus
EC	Ethyl cellulose
ECC	Emergency cardiac care; Extracorporeal circulation
ECCE	Extracapsular cataract extinction
ECD	Electron capture detector
ECG	Electrocardiogram
ECM	Electrochemical machining
ECMO	Extracorporeal membrane oxygenation
ECOD	Extracranial cerebrovascular occlusive disease
ECRI	Emergency Care Research Institute
ECS	Exner's Comprehensive System
ECT	Electroconvulsive shock therapy; Electroconvulsive therapy; Emission computed tomography
EDD	Estimated date of delivery
EDP	Aortic end diastolic pressure
EDTA	Ethylenediaminetetraacetic acid
EDX	Energy dispersive X-ray analysis
EEG	Electroencephalogram
EEI	Electrode electrolyte interface
EELV	End-expiratory lung volume
EER	Electrically evoked response
EF	Ejection fraction
EF	Electric field; Evoked magnetic fields
EFA	Estimated fetal age
EGF	Epidermal growth factor
EGG	Electrogastrogram
EIA	Enzyme immunoassay
EIU	Electrode impedance unbalance
ELF	Extra low frequency
ELGON	Electrical goniometer
ELISA	Enzyme-linked immunosorbent assay
ELS	Energy loss spectroscopy
ELV	Equivalent lung volume
EM	Electromagnetic
EMBS	Engineering in Medicine and Biology Society
emf	Electromotive force
EMG	Electromyogram
EMGE	Integrated electromyogram
EMI	Electromagnetic interference
EMS	Emergency medical services
EMT	Emergency medical technician
ENT	Ear, nose, and throat
EO	Elbow orthosis
EOG	Electrooculography
EOL	End of life
EOS	Eosinophil
EP	Elastoplastic; Evoked potentiate
EPA	Environmental protection agency
ER	Evoked response
ERCP	Endoscopic retrograde cholangiopancreatography
ERG	Electron radiography; Electroretinogram
ERMF	Event-related magnetic field
ERP	Event-related potential
ERV	Expiratory reserve volume

ESCA	Electron spectroscopy for chemical analysis
ESI	Electrode skin impedance
ESRD	End-stage renal disease
esu	Electrostatic unit
ESU	Electrosurgical unit
ESWL	Extracorporeal shock wave lithotripsy
ETO, Eto	Ethylene oxide
ETT	Exercise tolerance testing
EVA	Ethylene vinyl acetate
EVR	Endocardial viability ratio
EW	Extended wear
FAD	Flavin adenine dinucleotide
FARA	Flexible automation random analysis
FBD	Fetal biparietal diameter
FBS	Fetal bovine serum
fcc	Face centered cubic
FCC	Federal Communications Commission
Fct	Fluorocrit
FDA	Food and Drug Administration
FDCA	Food, Drug, and Cosmetic Act
FE	Finite element
FECG	Fetal electrocardiogram
FEF	Forced expiratory flow
FEL	Free electron lasers
FEM	Finite element method
FEP	Fluorinated ethylene propylene
FES	Functional electrical stimulation
FET	Field-effect transistor
FEV	Forced expiratory volume
FFD	Focal spot to film distance
FFT	Fast Fourier transform
FGF	Fresh gas flow
FHR	Fetal heart rate
FIC	Forced inspiratory capacity
FID	Flame ionization detector; Free-induction decay
FIFO	First-in-first-out
FITC	Fluorescent indicator tagged polymer
FL	Femur length
FM	Frequency modulation
FNS	Functional neuromuscular stimulation
FO	Foramen ovale
FO-CRT	Fiber optics cathode ray tube
FP	Fluorescence polarization
FPA	Fibrinopeptide A
FR	Federal Register
FRC	Federal Radiation Council; Functional residual capacity
FSD	Focus-to-surface distance
FTD	Focal spot to tissue-plane distance
FTIR	Fourier transform infrared
FTMS	Fourier transform mass spectrometer
FU	Fluorouracil
FUDR	Floxuridine
FVC	Forced vital capacity
FWHM	Full width at half maximum
FWTM	Full width at tenth maximum
GABA	Gamma amino buteric acid
GAG	Glycosaminoglycan
GBE	Gas-bearing electrodynamometer
GC	Gas chromatography; Guanine-cytosine
GDT	Gas discharge tube
GFR	Glomerular filtration rate
GHb	Glycosylated hemoglobin
GI	Gastrointestinal
GLC	Gas–liquid chromatography
GMV	General minimum variance
GNP	Gross national product
GPC	Giant papillary conjunctivitis
GPH	Gas-permeable hard
GPH-EW	Gas-permeable hard lens extended wear
GPO	Government Printing Office
GSC	Gas-solid chromatography
GSR	Galvanic skin response
GSWD	Generalized spike-wave discharge
HA	Hydroxyapatite
HAM	Helical axis of motion
Hb	Hemoglobin
HBE	His bundle electrogram
HBO	Hyperbaric oxygenation
HC	Head circumference
HCA	Hypothermic circulatory arrest
HCFA	Health care financing administration
HCL	Harvard Cyclotron Laboratory
hcp	Hexagonal close-packed
HCP	Half cell potential
HDPE	High density polyethylene
HECS	Hospital Equipment Control System
HEMS	Hospital Engineering Management System
HEPA	High efficiency particulate air filter
HES	Hydroxyethylstarch
HETP	Height equivalent to a theoretical plate
HF	High-frequency; Heating factor
HFCWO	High-frequency chest wall oscillation
HFER	High-frequency electromagnetic radiation
HFJV	High-frequency jet ventilation
HFO	High-frequency oscillator
HFOV	High-frequency oscillatory ventilation
HFPPV	High-frequency positive pressure ventilation
HFV	High-frequency ventilation
HHS	Department of Health and Human Services
HIBC	Health industry bar code
HIMA	Health Industry Manufacturers Association
HIP	Hydrostatic indifference point
HIS	Hospital information system
HK	Hexokinase
HL	Hearing level
HMBA	Hexamethylene bisacetamide
HMO	Health maintenance organization
HMWPE	High-molecular-weight polyethylene
HOL	Higher-order languages
HP	Heating factor; His-Purkinje
HpD	Hematoporphyrin derivative
HPLC	High-performance liquid chromatography
HPNS	High-pressure neurological syndrome
HPS	His-Purkinje system
HPX	High peroxidase activity

HR	Heart rate; High-resolution
HRNB	Halstead-Reitan Neuropsychological Battery
H/S	Hard/soft
HSA	Human serum albumin
HSG	Hysterosalpingogram
HTCA	Human tumor cloning assay
HTLV	Human T cell lymphotrophic virus
HU	Heat unit; Houndsfield units; Hydroxyurea
HVL	Half value layer
HVR	Hypoxic ventilatory response
HVT	Half-value thickness
IA	Image intensifier assembly; Inominate artery
IABP	Intraaortic balloon pumping
IAEA	International Atomic Energy Agency
IAIMS	Integrated Academic Information Management System
IASP	International Association for the Study of Pain
IC	Inspiratory capacity; Integrated circuit
ICCE	Intracapsular cataract extraction
ICD	Intracervical device
ICDA	International classification of diagnoses
ICL	Ms-clip lens
ICP	Inductively coupled plasma; Intracranial pressure
ICPA	Intracranial pressure amplitude
ICRP	International Commission on Radiological Protection
ICRU	International Commission on Radiological Units and Measurements
ICU	Intensive care unit
ID	Inside diameter
IDDM	Insulin dependent diabetes mellitus
IDE	Investigational device exemption
IDI	Index of inspired gas distribution
I:E	Inspiratory: expiratory
IEC	International Electrotechnical Commission; Ion-exchange chromatography
IEEE	Institute of Electrical and Electronics Engineers
IEP	Individual educational program
BETS	Inelastic electron tunneling spectroscopy
IF	Immunofluorescent
IFIP	International Federation for Information Processing
IFMBE	International Federation for Medical and Biological Engineering
IGFET	Insulated-gate field-effect transistor
IgG	Immunoglobulin G
IgM	Immunoglobulin M
IHP	Inner Helmholtz plane
IHSS	Idiopathic hypertrophic subaortic stenosis
II	Image intensifier
IIIES	Image intensifier input-exposure sensitivity
IM	Intramuscular
IMFET	Immunologically sensitive field-effect transistor
IMIA	International Medical Informatics Association
IMS	Information management system
IMV	Intermittent mandatory ventilation
INF	Interferon
IOL	Intraocular lens
IPC	Ion-pair chromatography
IPD	Intermittent peritoneal dialysis
IPG	Impedance plethysmography
IPI	Interpulse interval
IPPB	Intermittent positive pressure breathing
IPTS	International practical temperature scale
IR	Polyisoprene rubber
IRB	Institutional Review Board
IRBBB	Incomplete right bundle branch block
IRPA	International Radiation Protection Association
IRRAS	Infrared reflection-absorption spectroscopy
IRRS	Infrared reflection spectroscopy
IRS	Internal reflection spectroscopy
IRV	Inspiratory reserve capacity
IS	Image size; Ion-selective
ISC	Infant skin servo control
ISDA	Instantaneous screw displacement axis
ISE	Ion-selective electrode
ISFET	Ion-sensitive field effect transistor
ISIT	Intensified silicon-intensified target tube
ISO	International Organization for Standardization
ISS	Ion scattering spectroscopy
IT	Intrathecal
ITEP	Institute of Theoretical and Experimental Physics
ITEPI	Instantaneous trailing edge pulse impedance
ITLC	Instant thin-layer chromatography
IUD	Intrauterine device
IV	Intravenous
IVC	Inferior vena cava
IVP	Intraventricular pressure
JCAH	Joint Commission on the Accreditation of Hospitals
JND	Just noticeable difference
JRP	Joint replacement prosthesis
KB	Kent bundle
Kerma	Kinetic energy released in unit mass
KO	Knee orthosis
KPM	Kilopond meter
KRPB	Krebs-Ringer physiological buffer
LA	Left arm; Left atrium
LAD	Left anterior descending; Left axis deviation
LAE	Left atrial enlargement
LAK	Lymphokine activated killer
LAL	Limulus amoebocyte lysate
LAN	Local area network
LAP	Left atrial pressure
LAT	Left anterior temporalis
LBBB	Left bundle branch block
LC	Left carotid; Liquid chromatography

LCC	Left coronary cusp
LCD	Liquid crystal display
LDA	Laser Doppler anemometry
LDF	Laser Doppler flowmetry
LDH	Lactate dehydrogenase
LDPE	Low density polyethylene
LEBS	Low-energy brief stimulus
LED	Light-emitting diode
LEED	Low energy electron diffraction
LES	Lower esophageal sphincter
LESP	Lower esophageal sphincter pressure
LET	Linear energy transfer
LF	Low frequency
LH	Luteinizing hormone
LHT	Local hyperthermia
LL	Left leg
LLDPE	Linear low density polyethylene
LLPC	Liquid-liquid partition chromatography
LLW	Low-level waste
LM	Left masseter
LNNB	Luria-Nebraska Neuropsychological Battery
LOS	Length of stay
LP	Late potential; Lumboperitoneal
LPA	Left pulmonary artery
LPC	Linear predictive coding
LPT	Left posterior temporalis
LPV	Left pulmonary veins
LRP	Late receptor potential
LS	Left subclavian
LSC	Liquid-solid adsorption chromatography
LSI	Large scale integrated
LSV	Low-amplitude shear-wave viscoelastometry
LTI	Low temperature isotropic
LUC	Large unstained cells
LV	Left ventricle
LVAD	Left ventricular assist device
LVDT	Linear variable differential transformer
LVEP	Left ventricular ejection period
LVET	Left ventricular ejection time
LVH	Left ventricular hypertrophy
LYMPH	Lymphocyte
MAA	Macroaggregated albumin
MAC	Minimal auditory capabilities
MAN	Manubrium
MAP	Mean airway pressure; Mean arterial pressure
MAST	Military assistance to safety and traffic
MBA	Monoclonal antibody
MBV	Maximum breathing ventilation
MBX	Monitoring branch exchange
MCA	Methyl cryanoacrylate
MCG	Magnetocardiogram
MCI	Motion Control Incorporated
MCMI	Millon Clinical Multiaxial Inventory
MCT	Microcatheter transducer
MCV	Mean corpuscular volume
MDC	Medical diagnostic categories
MDI	Diphenylmethane diisocyanate; Medical Database Informatics
MDP	Mean diastolic aortic pressure
MDR	Medical device reporting
MDS	Multidimensional scaling
ME	Myoelectric
MED	Minimum erythema dose
MEDPAR	Medicare provider analysis and review
MEFV	Maximal expiratory flow volume
MEG	Magnetoencephalography
MeSH	Medline subject heading
METS	Metabolic equivalents
MF	Melamine-formaldehyde
MFP	Magnetic field potential
MGH	Massachusetts General Hospital
MHV	Magnetic heart vector
MI	Myocardial infarction
MIC	Minimum inhibitory concentration
MIFR	Maximum inspiratory flow rate
MINET	Medical Information Network
MIR	Mercury-in-rubber
MIS	Medical information system; Metal-insulator-semiconductor
MIT	Massachusetts Institute of Technology
MIT/BIH	Massachusetts Institute of Technology/ Beth Israel Hospital
MMA	Manual metal arc welding
MMA	Methyl methacrylate
MMECT	Multiple-monitored ECT
MMFR	Maximum midexpiratory flow rate
mm Hg	Millimeters of mercury
MMPI	Minnesota Multiphasic Personality Inventory
MMSE	Minimum mean square error
MO	Membrane oxygenation
MONO	Monocyte
MOSFET	Metal oxide silicon field-effect transistor
MP	Mercaptopurine; Metacarpal-phalangeal
MPD	Maximal permissible dose
MR	Magnetic resonance
MRG	Magnetoretinogram
MRI	Magnetic resonance imaging
MRS	Magnetic resonance spectroscopy
MRT	Mean residence time
MS	Mild steel; Multiple sclerosis
MSR	Magnetically shielded room
MTBF	Mean time between failure
MTF	Modulation transfer function
MTTR	Mean time to repair
MTX	Methotroxate
MUA	Motor unit activity
MUAP	Motor unit action potential
MUAPT	Motor unit action potential train
MUMPI	Missouri University Multi-Plane Imager
MUMPS	Massachusetts General Hospital utility multiuser programming system
MV	Mitral valve
MVO_2	Maximal oxygen uptake
MVTR	Moisture vapor transmission rate
MVV	Maximum voluntary ventilation
MW	Molecular weight

NAA	Neutron activation analysis
NAD	Nicotinamide adenine dinucleotide
NADH	Nicotinamide adenine dinucleotide, reduced form
NADP	Nicotinamide adenine dinucleotide phosphate
NAF	Neutrophil activating factor
NARM	Naturally occurring and accelerator-produced radioactive materials
NBB	Normal buffer base
NBD	Neuromuscular blocking drugs
N-BPC	Normal bonded phase chromatography
NBS	National Bureau of Standards
NCC	Noncoronary cusp
NCCLS	National Committee for Clinical Laboratory Standards; National Committee on Clinical Laboratory Standards
NCRP	National Council on Radiation Protection
NCT	Neutron capture theory
NEEP	Negative end-expiratory pressure
NEMA	National Electrical Manufacturers Association
NEMR	Nonionizing electromagnetic radiation
NEQ	Noise equivalent quanta
NET	Norethisterone
NEUT	Neutrophil
NFPA	National Fire Protection Association
NH	Neonatal hepatitis
NHE	Normal hydrogen electrode
NHLBI	National Heart, Lung, and Blood Institute
NIR	Nonionizing radiation
NIRS	National Institute for Radiologic Science
NK	Natural killer
NMJ	Neuromuscular junction
NMOS	N-type metal oxide silicon
NMR	Nuclear magnetic resonance
NMS	Neuromuscular stimulation
NPH	Normal pressure hydrocephalus
NPL	National Physical Laboratory
NR	Natural rubber
NRC	Nuclear Regulatory Commission
NRZ	Non-return-to-zero
NTC	Negative temperature coefficient
NTIS	National Technical Information Service
NVT	Neutrons versus time
NYHA	New York Heart Association
ob/gyn	Obstetrics and gynecology
OCR	Off-center ratio; Optical character recognition
OCV	Open circuit voltage
OD	Optical density; Outside diameter
ODC	Oxyhemoglobin dissociation curve
ODT	Oxygen delivery truck
ODU	Optical density unit
OER	Oxygen enhancement ratio
OFD	Object to film distance; Occiputo-frontal diameter
OHL	Outer Helmholtz layer
OHP	Outer Helmholtz plane
OIH	Orthoiodohippurate
OPG	Ocular pneumoplethysmography
OR	Operating room
OS	Object of known size; Operating system
OTC	Over the counter
OV	Offset voltage
PA	Posterioanterior; Pulmonary artery; Pulse amplitude
PACS	Picture archiving and communications systems
PAD	Primary afferent depolarization
PAM	Pulse amplitude modulation
PAN	Polyacrylonitrile
PAP	Pulmonary artery pressure
PAR	Photoactivation ratio
PARFR	Program for Applied Research on Fertility Regulation
PARR	Poetanesthesia recovery room
PAS	Photoacoustic spectroscopy
PASG	Pneumatic antishock garment
PBI	Penile brachial index
PBL	Positive beam limitation
PBT	Polybutylene terephthalate
PC	Paper chromatography; Personal computer; Polycarbonate
PCA	Patient controlled analgesia; Principal components factor analysis
PCG	Phonocardiogram
PCI	Physiological cost index
PCL	Polycaprolactone; Posterior chamber lens
PCR	Percent regurgitation
PCRC	Perinatal Clinical Research Center
PCS	Patient care system
PCT	Porphyria cutanea tarda
PCWP	Pulmonary capillary wedge pressure
PD	Peritoneal dialysis; Poly-p-dioxanone; Potential difference; Proportional and derivative
PDD	Percent depth dose; Perinatal Data Directory
PDE	Pregelled disposable electrodes
p.d.f.	Probability density function
PDL	Periodontal ligament
PDM	Pulse duration modulation
PDMSX	Polydimethyl siloxane
PDS	Polydioxanone
PE	Polyethylene
PEEP	Positive end-expiratory pressure
PEFR	Peak expiratory now rate
PEN	Parenteral and enteral nutrition
PEP	Preejection period
PEPPER	Programs examine phonetic find phonological evaluation records
PET	Polyethylene terephthalate; Positron-emission tomography
PEU	Polyetherurethane
PF	Platelet factor
PFA	Phosphonoformic add
PFC	Petrofluorochemical
PFT	Pulmonary function testing
PG	Polyglycolide; Propylene glycol

PGA	Polyglycolic add
PHA	Phytohemagglutinin; Pulse-height analyzer
PHEMA	Poly-2-hydroxyethyl methacrylate
PI	Propidium iodide
PID	Pelvic inflammatory disease; Proportional/integral/derivative
PIP	Peak inspiratory pressure
PL	Posterior leaflet
PLA	Polylactic acid
PLATO	Program Logic for Automated Teaching Operations
PLD	Potentially lethal damage
PLED	Periodic latoralized epileptiform discharge
PLT	Platelet
PM	Papillary muscles; Preventive maintenance
PMA	Polymethyl acrylate
p.m.f.	Probability mass function
PMMA	Polymethyl methacrylate
PMOS	P-type metal oxide silicon
PMP	Patient management problem; Poly(4-methylpentane)
PMT	Photomultiplier tube
PO	Per os
Po_2	Partial pressure of oxygen
POBT	Polyoxybutylene terephthalate
POM	Polyoxymethylene
POMC	Patient order management and communication system
POPRAS	Problem Oriented Perinatal Risk Assessment System
PP	Perfusion pressure; Polyproplyene; Postprandial (after meals)
PPA	Phonemic process analysis
PPF	Plasma protein fraction
PPM	Pulse position modulation
PPSFH	Polymerized phyridoxalated stroma-free hemoglobin
PR	Pattern recognition; Pulse rate
PRBS	Pseudo-random binary signals
PRP	Pulse repetition frequency
PRO	Professional review organization
PROM	Programmable read only memory
PS	Polystyrene
PSA	Pressure-sensitive adhesive
PSF	Point spread function
PSI	Primary skin irritation
PSP	Postsynaptic potential
PSR	Proton spin resonance
PSS	Progressive systemic sclerosis
PT	Plasma thromboplastin
PTB	Patellar tendon bearing orthosis
PTC	Plasma thromboplastin component; Positive temperature coefficient; Pressurized personal transfer capsule
PTCA	Percutaneous transluminal coronary angioplasty
PTFE	Polytetrafluoroethylene
PTT	Partial thromboplastin time
PUL	Percutaneous ultrasonic lithotripsy
PURA	Prolonged ultraviolet-A radiation
PUVA	Psoralens and longwave ultraviolet light photochemotherapy
P/V	Pressure/volume
PVC	Polyvinyl chloride; Premature ventricular contraction
PVI	Pressure–volume index
PW	Pulse wave; Pulse width
PWM	Pulse width modulation
PXE	Pseudo-xanthoma elasticum
QA	Quality assurance
QC	Quality control
R-BPC	Reverse bonded phase chromatography
R/S	Radiopaque-spherical
RA	Respiratory amplitude; Right arm
RAD	Right axis deviation
RAE	Right atrial enlargement
RAM	Random access memory
RAP	Right atrial pressure
RAT	Right anterior temporalis
RB	Right bundle
RBBB	Right bundle branch block
RBC	Red blood cell
RBE	Relative biologic effectiveness
RBF	Rose bengal fecal excretion
RBI	Resting baseline impedance
RCBD	Randomized complete block diagram
rCBF	Regional cerebral blood flow
RCC	Right coronary cusp
RCE	Resistive contact electrode
R&D	Research and development
r.e.	Random experiment
RE	Reference electrode
REM	Rapid eye movement; Return electrode monitor
REMATE	Remote access and telecommunication system
RES	Reticuloendothelial system
RESNA	Rehabilitation Engineering Society of North America
RF	Radio frequency; Radiographic-nuoroscopic
RFI	Radio-frequency interference
RFP	Request for proposal
RFQ	Request for quotation
RH	Relative humidity
RHE	Reversible hydrogen electrode
RIA	Radioimmunoassay
RM	Repetition maximum; Right masseter
RMR	Resting metabolic rate
RMS	Root mean square
RN	Radionuclide
RNCA	Radionuclide cineagiogram
ROI	Regions of interest
ROM	Range of motion; Read only memory
RP	Retinitis pigmentosa
RPA	Right pulmonary artery
RPP	Rate pressure product
RPT	Rapid pull-through technique
RPV	Right pulmonary veins
RQ	Respiratory quotient

RR	Recovery room
RRT	Recovery room time; Right posterior temporalis
RT	Reaction time
RTD	Resistance temperature device
RTT	Revised token test
r.v.	Random variable
RV	Residual volume; Right ventricle
RVH	Right ventricular hypertrophy
RVOT	Right ventricular outflow tract
RZ	Return-to-zero
SA	Sinoatrial; Specific absorption
SACH	Solid-ankle-cushion-heel
SAD	Source-axis distance; Statistical Analysis System
SAINT	System analysis of integrated network of tasks
SAL	Sterility assurance level; Surface averaged lead
SALT	Systematic analysis of language transcripts
SAMI	Socially acceptable monitoring instrument
SAP	Systemic arterial pressure
SAR	Scatter-air ratio; Specific absorption rate
SARA	System for anesthetic and respiratory gas analysis
SBE	Subbacterial endocarditis
SBR	Styrene-butadiene rubbers
SC	Stratum corneum; Subcommittees
SCAP	Right scapula
SCE	Saturated calomel electrode; Sister chromatid exchange
SCI	Spinal cord injury
SCRAD	Sub-Committee on Radiation Dosimetry
SCS	Spinal cord stimulation
SCUBA	Self-contained underwater breathing apparatus
SD	Standard deviation
SDA	Stepwise discriminant analysis
SDS	Sodium dodecyl sulfate
S&E	Safety and effectiveness
SE	Standard error
SEC	Size exclusion chromatography
SEM	Scanning electron microscope; Standard error of the mean
SEP	Somatosensory evoked potential
SEXAFS	Surface extended X-ray absorption fine structure
SF	Surviving fraction
SFD	Source-film distance
SFH	Stroma-free hemoglobin
SFTR	Sagittal frontal transverse rotational
SG	Silica gel
SGF	Silica gel fraction
SGG	Spark gap generator
SGOT	Serum glutamic oxaloacetic transaminase
SGP	Strain gage plethysmography; Stress-generated potential
SHE	Standard hydrogen electrode
SI	Le Système International d'Unités
SEBS	Surgical isolation barrier system
SID	Source to image reception distance
SIMFU	Scanned intensity modulated focused ultrasound
SIMS	Secondary ion mass spectroscopy; System for isometric muscle strength
SISI	Short increment sensitivity index
SL	Surgical lithotomy
SLD	Sublethal damage
SLE	Systemic lupus erythemotodes
SMA	Sequential multiple analyzer
SMAC	Sequential multiple analyzer with computer
SMR	Sensorimotor
S/N	Signal-to-noise
S:N/D	Signal-to-noise ratio per unit dose
SNP	Sodium nitroprusside
SNR	Signal-to-noise ratio
SOA	Sources of artifact
SOAP	Subjective, objective, assessment, plan
SOBP	Spread-out Bragg peak
SP	Skin potential
SPECT	Single photon emission computed tomography
SPL	Sound pressure level
SPRINT	Single photon ring tomograph
SPRT	Standard platinum resistance thermometer
SPSS	Statistical Package for the Social Sciences
SQUID	Superconducting quantum interference device
SQV	Square wave voltammetry
SR	Polysulfide rubbers
SRT	Speech reception threshold
SS	Stainless steel
SSB	Single strand breaks
SSD	Source-to-skin distance; Source-to-surface distance
SSE	Stainless steel electrode
SSEP	Somatosensory evoked potential
SSG	Solid state generator
SSP	Skin stretch potential
SSS	Sick sinus syndrome
STD	Source-tray distance
STI	Systolic time intervals
STP	Standard temperature and pressure
STPD	Standard temperature pressure dry
SV	Stroke volume
SVC	Superior vena cava
SW	Standing wave
TAA	Tumor-associated antigens
TAC	Time-averaged concentration
TAD	Transverse abdominal diameter
TAG	Technical Advisory Group
TAH	Total artificial heart
TAR	Tissue-air ratio
TC	Technical Committees
TCA	Tricarboxylic acid cycle
TCD	Thermal conductivity detector
TCES	Transcutaneous cranial electrical stimulation

TCP	Tricalcium phosphate
TDD	Telecommunication devices for the deaf
TDM	Therapeutic drug monitoring
TE	Test electrode; Thermoplastic elastomers
TEAM	Technology evaluation and acquisition methods
TEM	Transmission electron microscope; Transverse electric and magnetic mode; Transverse electromagnetic mode
TENS	Transcutaneous electrical nerve stimulation
TEP	Tracheoesophageal puncture
TEPA	Triethylenepho-sphoramide
TF	Transmission factor
TFE	Tetrafluorethylene
TI	Totally implantable
TICCIT	Time-shared Interaction Computer-Controlled Information Television
TLC	Thin-layer chromatography; Total lung capacity
TLD	Thermoluminescent dosimetry
TMJ	Temporomandibular joint
TMR	Tissue maximum ratio; Topical magnetic resonance
TNF	Tumor necrosis factor
TOF	Train-of-four
TP	Thermal performance
TPC	Temperature pressure correction
TPD	Triphasic dissociation
TPG	Transvalvular pressure gradient
TPN	Total parenteral nutrition
TR	Temperature rise
tRNA	Transfer RNA
TSH	Thyroid stimulating hormone
TSS	Toxic shock syndrome
TTD	Telephone devices for the deaf
TTI	Tension time index
TTR	Transition temperature range
TTV	Trimming tip version
TTY	Teletypewriter
TUR	Transurethral resection
TURP	Transurethral resections of the prostrate
TV	Television; Tidal volume; Tricuspid valve
TVER	Transscleral visual evoked response
TW	Traveling wave
TxB_2	Thrombozame B^2
TZ	Transformation zone
UES	Upper esophageal sphincter
UP	Urea-formaldehyde
UfflS	University Hospital Information System
UHMW	Ultra high molecular weight
UHMWPE	Ultra high molecular weight polyethylene
UL	Underwriters Laboratory
ULF	Ultralow frequency
ULTI	Ultralow temperature isotropic
UMN	Upper motor neuron
UO	Urinary output
UPTD	Unit pulmonary oxygen toxicity doses
UR	Unconditioned response
US	Ultrasound; Unconditioned stimulus
USNC	United States National Committee
USP	United States Pharmacopeia
UTS	Ultimate tensile strength
UV	Ultraviolet; Umbilical vessel
UVR	Ultraviolet radiation
V/F	Voltage-to-frequency
VA	Veterans Administration
VAS	Visual analog scale
VBA	Vaginal blood volume in arousal
VC	Vital capacity
VCO	Voltage-controlled oscillator
VDT	Video display terminal
VECG	Vectorelectrocardiography
VEP	Visually evoked potential
VF	Ventricular fibrillation
VOP	Venous occlusion plethysmography
VP	Ventriculoperitoneal
VPA	Vaginal pressure pulse in arousal
VPB	Ventricular premature beat
VPR	Volume pressure response
VSD	Ventricular septal defect
VSWR	Voltage standing wave ratio
VT	Ventricular tachycardia
VTG	Vacuum tube generator
VTS	Viewscan text system
VV	Variable version
WAIS-R	Weschler Adult Intelligence Scale-Revised
WAK	Wearable artificial kidney
WAML	Wide-angle mobility light
WBAR	Whole-body autoradiography
WBC	White blood cell
WG	Working Groups
WHO	World Health Organization; Wrist hand orthosis
WLF	Williams-Landel-Ferry
WMR	Work metabolic rate
w/o	Weight percent
WORM	Write once, read many
WPW	Wolff-Parkinson-White
XPS	X-ray photon spectroscopy
XR	Xeroradiograph
YAG	Yttrium aluminum garnet
ZPL	Zero pressure level

CONVERSION FACTORS AND UNIT SYMBOLS

SI UNITS (ADOPTED 1960)

A new system of metric measurement, the International System of Units (abbreviated SI), is being implemented throughout the world. This system is a modernized version of the MKSA (meter, kilogram, second, ampere) system, and its details are published and controlled by an international treaty organization (The International Bureau of Weights and Measures).

SI units are divided into three classes:

Base Units

length	meter[†] (m)
mass[‡]	kilogram (kg)
time	second (s)
electric current	ampere (A)
thermodynamic temperature§	kelvin (K)
amount of substance	mole (mol)
luminous intensity	candela (cd)

Supplementary Units

plane angle	radian (rad)
solid angle	steradian (sr)

Derived Units and Other Acceptable Units

These units are formed by combining base units, supplementary units, and other derived units. Those derived units having special names and symbols are marked with an asterisk (*) in the list below:

Quantity	*Unit*	*Symbol*	*Acceptable equivalent*
*absorbed dose	gray	Gy	J/kg
acceleration	meter per second squared	m/s^2	
*activity (of ionizing radiation source)	becquerel	Bq	1/s
area	square kilometer	km^2	
	square hectometer	hm^2	ha (hectare)
	square meter	m^2	

[†]The spellings "metre" and "litre" are preferred by American Society for Testing and Materials (ASTM); however, "−er" will be used in the Encyclopedia.

[‡]"Weight" is the commonly used term for "mass."

§Wide use is made of "Celsius temperature" (t) defined $t = T - T_0$ where T is the thermodynamic temperature, expressed in kelvins, and $T_0 = 273.15\,K$ by definition. A temperature interval may be expressed in degrees Celsius as well as in kelvins.

Quantity	*Unit*	*Symbol*	*Acceptable equivalent*
*capacitance	farad	F	C/V
concentration (of amount of substance)	mole per cubic meter	mol/m^3	
*conductance	siemens	S	A/V
current density	ampere per square meter	A/m^2	
density, mass density	kilogram per cubic meter	kg/m^3	g/L; mg/cm^3
dipole moment (quantity)	coulomb meter	C·m	
*electric charge, quantity of electricity	coulomb	C	A·s
electric charge density	coulomb per cubic meter	C/m^3	
electric field strength	volt per meter	V/m	
electric flux density	coulomb per square meter	C/m^2	
*electric potential, potential difference, electromotive force	volt	V	W/A
*electric resistance	ohm	Ω	V/A
*energy, work, quantity of heat	megajoule	MJ	
	kilojoule	kJ	
	joule	J	N·m
	electron volt†	eV†	
	kilowatt hour†	kW·h†	
energy density	joule per cubic meter	J/m^3	
*force	kilonewton	kN	
	newton	N	$kg·m/s^2$
*frequency	megahertz	MHz	
	hertz	Hz	1/s
heat capacity, entropy	joule per kelvin	J/K	
heat capacity (specific), specific entropy	joule per kilogram kelvin	J/(kg·K)	
heat transfer coefficient	watt per square meter kelvin	$W/(m^2·K)$	
*illuminance	lux	lx	lm/m^2
*inductance	henry	H	Wb/A
linear density	kilogram per meter	kg/m	
luminance	candela per square meter	cd/m^2	
*luminous flux	lumen	lm	cd·sr
magnetic field strength	ampere per meter	A/m	
*magnetic flux	weber	Wb	V·s
*magnetic flux density	tesla	T	Wb/m^2
molar energy	joule per mole	J/mol	
molar entropy, molar heat capacity	joule per mole kelvin	J/(mol·K)	
moment of force, torque	newton meter	N·m	
momentum	kilogram meter per second	kg·m/s	
permeability	henry per meter	H/m	
permittivity	farad per meter	F/m	
*power, heat flow rate, radiant flux	kilowatt	kW	
	watt	W	J/s
power density, heat flux density, irradiance	watt per square meter	W/m^2	
*pressure, stress	megapascal	MPa	
	kilopascal	kPa	
	pascal	Pa	N/m^2
sound level	decibel	dB	
specific energy	joule per kilogram	J/kg	
specific volume	cubic meter per kilogram	m^3/kg	
surface tension	newton per meter	N/m	
thermal conductivity	watt per meter kelvin	W/(m·K)	
velocity	meter per second	m/s	
	kilometer per hour	km/h	
viscosity, dynamic	pascal second	Pa·s	
	millipascal second	mPa·s	

†This non-SI unit is recognized as having to be retained because of practical importance or use in specialized fields.

Quantity	*Unit*	*Symbol*	*Acceptable equivalent*
viscosity, kinematic	square meter per second	m^2/s	
	square millimeter per second	mm^2/s	
	cubic meter	m^3	
	cubic decimeter	dm^3	L(liter)
	cubic centimeter	cm^3	mL
wave number	1 per meter	m^{-1}	
	1 per centimeter	cm^{-1}	

In addition, there are 16 prefixes used to indicate order of magnitude, as follows:

Multiplication factor	*Prefix*	*Symbol*	*Note*
10^{18}	exa	E	
10^{15}	peta	P	
10^{12}	tera	T	
10^{9}	giga	G	
10^{8}	mega	M	
10^{3}	kilo	k	
10^{2}	hecto	h[a]	[a]Although hecto, deka, deci, and centi are
10	deka	da[a]	SI prefixes, their use should be avoided
10^{-1}	deci	d[a]	except for SI unit-multiples for area and
10^{-2}	centi	c[a]	volume and nontechnical use of
10^{-3}	milli	m	centimeter, as for body and clothing
10^{-6}	micro	μ	measurement.
10^{-9}	nano	n	
10^{-12}	pico	p	
10^{-15}	femto	f	
10^{-18}	atto	a	

For a complete description of SI and its use the reader is referred to ASTM E 380.

CONVERSION FACTORS TO SI UNITS

A representative list of conversion factors from non-SI to SI units is presented herewith. Factors are given to four significant figures. Exact relationships are followed by a dagger (†). A more complete list is given in ASTM E 380-76 and ANSI Z210. 1-1976.

To convert from	*To*	*Multiply by*
acre	square meter (m^2)	4.047×10^{3}
angstrom	meter (m)	$1.0 \times 10^{-10\dagger}$
are	square meter (m^2)	$1.0 \times 10^{2\dagger}$
astronomical unit	meter (m)	1.496×10^{11}
atmosphere	pascal (Pa)	1.013×10^{5}
bar	pascal (Pa)	$1.0 \times 10^{5\dagger}$
barrel (42 U.S. liquid gallons)	cubic meter (m^3)	0.1590
Btu (International Table)	joule (J)	1.055×10^{3}
Btu (mean)	joule (J)	1.056×10^{3}
Bt (thermochemical)	joule (J)	1.054×10^{3}
bushel	cubic meter (m^3)	3.524×10^{-2}
calorie (International Table)	joule (J)	4.187
calorie (mean)	joule (J)	4.190
calorie (thermochemical)	joule (J)	$4.184^{\dagger}$
centimeters of water (39.2 °F)	pascal (Pa)	98.07
centipoise	pascal second (Pa·s)	$1.0 \times 10^{-3\dagger}$
centistokes	square millimeter per second (mm^2/s)	$1.0^{\dagger}$

To convert from	*To*	*Multiply by*
cfm (cubic foot per minute)	cubic meter per second (m^3/s)	4.72×10^{-4}
cubic inch	cubic meter (m^3)	1.639×10^{-4}
cubic foot	cubic meter (m^3)	2.832×10^{-2}
cubic yard	cubic meter (m^3)	0.7646
curie	becquerel (Bq)	$3.70 \times 10^{10\dagger}$
debye	coulomb-meter (C·m)	3.336×10^{-30}
degree (angle)	radian (rad)	1.745×10^{-2}
denier (international)	kilogram per meter (kg/m)	1.111×10^{-7}
	tex	0.1111
dram (apothecaries')	kilogram (kg)	3.888×10^{-3}
dram (avoirdupois)	kilogram (kg)	1.772×10^{-3}
dram (U.S. fluid)	cubic meter (m^3)	3.697×10^{-6}
dyne	newton(N)	$1.0 \times 10^{-6\dagger}$
dyne/cm	newton per meter (N/m)	$1.00 \times 10^{-3\dagger}$
electron volt	joule (J)	1.602×10^{-19}
erg	joule (J)	$1.0 \times 10^{-7\dagger}$
fathom	meter (m)	1.829
fluid ounce (U.S.)	cubic meter (m^3)	2.957×10^{-5}
foot	meter (m)	$0.3048^{\dagger}$
foot-pound force	joule (J)	1.356
foot-pound force	newton meter (N·m)	1.356
foot-pound force per second	watt(W)	1.356
footcandle	lux (lx)	10.76
furlong	meter (m)	2.012×10^2
gal	meter per second squared (m/s^2)	$1.0 \times 10^{-2\dagger}$
gallon (U.S. dry)	cubic meter (m^3)	4.405×10^{-3}
gallon (U.S. liquid)	cubic meter (m^3)	3.785×10^{-3}
gilbert	ampere (A)	0.7958
gill (U.S.)	cubic meter (m^3)	1.183×10^{-4}
grad	radian	1.571×10^{-2}
grain	kilogram (kg)	6.480×10^{-5}
gram force per denier	newton per tex (N/tex)	8.826×10^{-2}
hectare	square meter (m^2)	$1.0 \times 10^{4\dagger}$
horsepower (550 ft·lbf/s)	watt(W)	7.457×10^2
horsepower (boiler)	watt(W)	9.810×10^3
horsepower (electric)	watt(W)	$7.46 \times 10^{2\dagger}$
hundredweight (long)	kilogram (kg)	50.80
hundredweight (short)	kilogram (kg)	45.36
inch	meter (m)	$2.54 \times 10^{-2\dagger}$
inch of mercury (32 °F)	pascal (Pa)	3.386×10^3
inch of water (39.2 °F)	pascal (Pa)	2.491×10^2
kilogram force	newton (N)	9.807
kilopond	newton (N)	9.807
kilopond-meter	newton-meter (N·m)	9.807
kilopond-meter per second	watt (W)	9.807
kilopond-meter per min	watt(W)	0.1635
kilowatt hour	megajoule (MJ)	$3.6^{\dagger}$
kip	newton (N)	4.448×10^2
knot international	meter per second (m/s)	0.5144
lambert	candela per square meter (cd/m^2)	3.183×10^3
league (British nautical)	meter (m)	5.559×10^2
league (statute)	meter (m)	4.828×10^3
light year	meter (m)	9.461×10^{15}
liter (for fluids only)	cubic meter (m^3)	$1.0 \times 10^{-3\dagger}$
maxwell	weber (Wb)	$1.0 \times 10^{-8\dagger}$
micron	meter (m)	$1.0 \times 10^{-6\dagger}$
mil	meter (m)	$2.54 \times 10^{-5\dagger}$
mile (U.S. nautical)	meter (m)	$1.852 \times 10^{3\dagger}$
mile (statute)	meter (m)	1.609×10^3
mile per hour	meter per second (m/s)	0.4470

To convert from	*To*	*Multiply by*
millibar	pascal (Pa)	1.0×10^{2}
millimeter of mercury (0 °C)	pascal (Pa)	1.333×10^{2}†
millimeter of water (39.2 °F)	pascal (Pa)	9.807
minute (angular)	radian	2.909×10^{-4}
myriagram	kilogram (kg)	10
myriameter	kilometer (km)	10
oersted	ampere per meter (A/m)	79.58
ounce (avoirdupois)	kilogram (kg)	2.835×10^{-2}
ounce (troy)	kilogram (kg)	3.110×10^{-2}
ounce (U.S. fluid)	cubic meter (m^3)	2.957×10^{-5}
ounce-force	newton (N)	0.2780
peck (U.S.)	cubic meter (m^3)	8.810×10^{-3}
pennyweight	kilogram (kg)	1.555×10^{-3}
pint (U.S. dry)	cubic meter (m^3)	5.506×10^{-4}
pint (U.S. liquid)	cubic meter (m^3)	4.732×10^{-4}
poise (absolute viscosity)	pascal second (Pa·s)	0.10†
pound (avoirdupois)	kilogram (kg)	0.4536
pound (troy)	kilogram (kg)	0.3732
poundal	newton (N)	0.1383
pound-force	newton (N)	4.448
pound per square inch (psi)	pascal (Pa)	6.895×10^{3}
quart (U.S. dry)	cubic meter (m^3)	1.101×10^{-3}
quart (U.S. liquid)	cubic meter (m^3)	9.464×10^{-4}
quintal	kilogram (kg)	1.0×10^{2}†
rad	gray (Gy)	1.0×10^{-2}†
rod	meter (m)	5.029
roentgen	coulomb per kilogram (C/kg)	2.58×10^{-4}
second (angle)	radian (rad)	4.848×10^{-6}
section	square meter (m^2)	2.590×10^{6}
slug	kilogram (kg)	14.59
spherical candle power	lumen (lm)	12.57
square inch	square meter (m^2)	6.452×10^{-4}
square foot	square meter (m^2)	9.290×10^{-2}
square mile	square meter (m^2)	2.590×10^{6}
square yard	square meter (m^2)	0.8361
store	cubic meter (m^3)	1.0†
stokes (kinematic viscosity)	square meter per second (m^2/s)	1.0×10^{-4}†
tex	kilogram per meter (kg/m)	1.0×10^{-6}†
ton (long, 2240 pounds)	kilogram (kg)	1.016×10^{3}
ton (metric)	kilogram (kg)	1.0×10^{3}†
ton (short, 2000 pounds)	kilogram (kg)	9.072×10^{2}
torr	pascal (Pa)	1.333×10^{2}
unit pole	weber (Wb)	1.257×10^{-7}
yard	meter (m)	0.9144†

A

ABLATION. See TISSUE ABLATION.

ABSORBABLE BIOMATERIALS. See BIOMATERIALS, ABSORBABLE.

ACRYLIC BONE CEMENT. See BONE CEMENT, ACRYLIC.

ACTINOTHERAPY. See ULTRAVIOLET RADIATION IN MEDICINE.

ADOPTIVE IMMUNOTHERAPY. See IMMUNOTHERAPY.

AFFINITY CHROMATOGRAPHY. See CHROMATOGRAPHY.

ALLOYS, SHAPE MEMORY

YOUNG KON KIM
Inje University
Kimhae City
Korea

INTRODUCTION

An alloy is defined as a substance with metallic properties that is composed of two or more chemical elements of which at least one is an elemental metal (1). The internal structure of most alloys starts to change only when it is no longer stable. When external influences, such as pressure and temperature, are varied, it will tend to transform spontaneously into a mixture of phases, the structures, compositions, and morphologies of which differ from the initial one. Such microstructural changes are known as phase transformation and may involve considerable atomic rearrangement and compositional change (2,3).

Shape memory alloys (SMAs) exhibit a unique mechanical "memory", or restoration force characteristic, when heated above a certain phase-transformation temperature range (TTR), after having been deformed below the TTR. This thermally activated shape recovering behavior is called the shape memory effect (SME) (3–5). This particular effect is closely related to a martensitic phase transformation accompanied by subatomic shear deformation resulting from the diffusionless, cooperative movement of atoms (6,7). The name martensite was originally used to describe the very fine, hard microstructure found in quenched steels (8). The meaning of this word has been extended gradually to describe the microstructure of nonferrous alloys that have similar characteristics.

SMAs have two stable phases: a high temperature stable phase, called the parent or austenite phase and a low temperature stable martensite phase. Martensite phases can be induced by cooling or stressing and are called thermally induced martensite (TIM) or stress induced martensite (SIM), respectively (8). The TIM forms and grows continuously as the temperature is lowered, and it shrinks and vanishes as the temperature is raised. The SIM is generated continuously with increasing applied stress on the alloy. On removing the applied stress, SIM disappears gradually at a constant temperature. If the temperature is sufficiently low when stressing, however, the SIM cannot return to its initial structure when the stress is removed. When the temperature is increased above the TTR, the residual SIM restores the original structure, resulting in shape recovery (9). Surprisingly, this process can be reliably repeated millions of times, provided that the strain limits are not breached. If dislocations or slips intervene in this process, the shape memory becomes imperfect. When the applied stress on a SMA is too great, irreversible slip occurs, and the SMA cannot recover its original shape even after heating above TTR (10). However, it can remember this hot parent pattern. In the next cooling cycle, the SMA changes slightly and remembers the cool-martensite pattern. A SMA trained with this repeated cyclic treatment is called a two-way SMA (9). A schematic explanation of the SME related to the two-dimensional (2D) crystal structure (11) is shown in Fig. 1. When a SMA is cooled below its TTR, the parent phase begins to form TIM without an external shape change. This TIM can be changed into SIM easily by mechanical deformation below the TTR. When the deformed SMA is heated above its TTR, however, it cannot hold the deformed shape anymore, and the SMA returns to its original shape, resulting in a reverse martensitic phase transformation.

A SMA also shows rubber-like behavior at temperatures above its TTR. When a SMA is deformed isothermally above its TTR, only SIM is produced, until plastic deformation occurs. Then, the SIM disappears immediately after removing the applied load, resulting in a much greater amount of recovering strain, in excess of the elastic limit, compared to the conventional elastic strain of a metal. This rubber-like behavior at a constant temperature above TTR is called superelasticity (12). A schematic explanation of superelasticity is shown in Fig. 2.

These contrasting behaviors of superelasticity and SME are a function of the testing temperature. If a SMA is tested below its TTR, it shows SME, while a SMA that is deformed above its TTR shows superelasticity.

It is convenient to subdivide the superelastic behavior into two categories, "superelasticity" and "rubber-like behavior", depending on the nature of the driving forces and mechanism involved. If it is triggered by SIM formation and subsequent reversion, the terminology superelasticity is used. By contrast, rubber-like behavior does not involve phase transformation, but involves deformation of the martensite itself. It is closely related to the reversible movement of deformed twin boundaries or martensite boundaries (10).

An example of SME in a shape-memory suture needle (13) is shown in Fig. 3. Figure 3a shows a curved needle with the shape preset by a heat-treatment process. When the shape-memory needle is cooled below its TTR, it is readily amenable to a change in shape with forceps (b). On heating it above TTR, thermal energy causes the needle to recover its original curved shape (c).

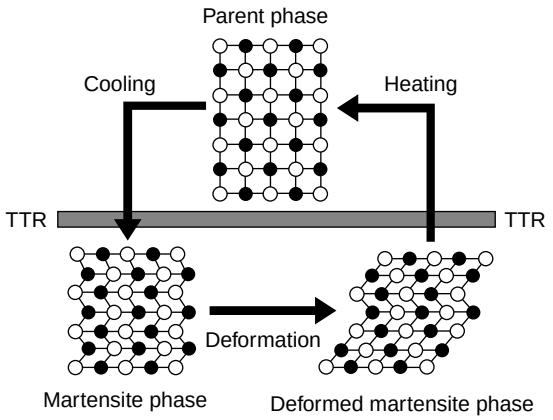


Figure 1. Schematic illustration of the shape memory effect. The parent phase is cooled below TTR to form a twinned (self-accommodated) martensite without an external shape change. Deformed martensite is produced with twin boundary movement and a change of shape by deformation below the TTR. Heating above the TTR results in reverse transformation and leads to shape recovery.

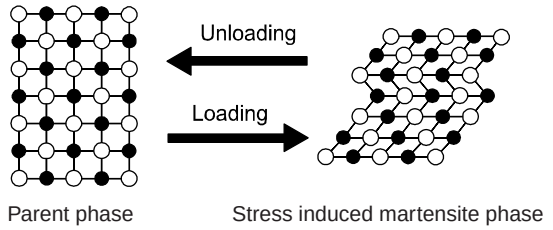


Figure 2. Schematic illustration of the superelasticity of a SMA above TTR. During the loading process, the applied load changes the parent phase into stress-induced martensite, which disappears instantly on unloading.

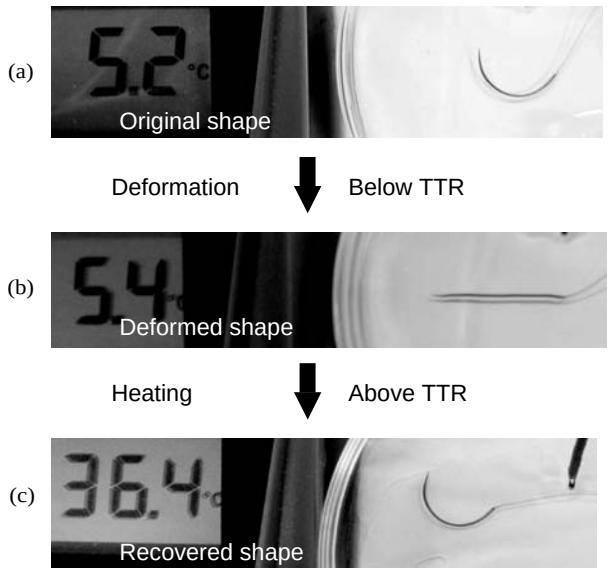


Figure 3. Shape-memory effect in a SMA suture needle. (a) Cooling the SMA suture needle below its TTR, (b) straightening the SMA suture needle below its TTR, (c) recovering the original shape of the SMA suture needle above its TTR.

HISTORY OF SHAPE MEMORY ALLOYS

The first observed shape memory phenomenon was pseudoelasticity. In 1932, Oelander observed it in a Au–Cd alloy and called it "rubber-like" behavior (14). Owing to the great amount of reversible strain, this effect is also called "superelasticity". The SME was discovered in 1938 by Greninger and Mooradian (15), while observing the formation and disappearance of martensite with falling and rising temperature in a brass (Cu–Zn alloy) sample. The maximum amount of reversible strain was observed in a Cu–Al–Ni single crystal with a recoverable elastic strain of 24% (16). In 1949, Kurdjumov and Khandros (17), provided a theoretical explanation of the basic mechanism of SME, the thermoelastic behavior of the martensite phase in Au–Cd alloy. Numerous alloy systems have been found to exhibit shape memory behavior. However, the great breakthrough came in 1963, when Buehler et al. (4) at the U.S. Naval Ordnance Laboratory discovered the SME in an equiatomic alloy of Ni–Ti, since then popularized under the name nitinol (Nickel–Titanium Naval Ordnance Laboratory). Partial listings of SMAs include the alloy systems: Ag–Cd, Au–Cd, Au–Cu, Cu–Zn, Cu–Zn–X (X = Si, Sn, Al, Ga), Cu–Al–Ni, Cu–Au–Zn, Cu–Sn, Ni–Al, Ni–Nb, Ni–Ti, Ni–Ti–Cu, Ti–Pd–Ni, In–Tl, In–Cd, Mn–Cd, Fe–Ni, Fe–Mn, Fe–Pt, Fe–Pd, and Fe–Ni–Co–Ti (9). It took several years to understand the microscopic, crystallographic, and thermodynamic properties of these extraordinary metals (18–20). The aeronautical, mechanical, electrical, biomedical, and biological engineering communities, as well as the health professions, are making use of shape memory alloys for a wide range of applications (9). Several commercial applications of Ni–Ti and Cu–Zn–Al SMAs have been developed, such as tube-fitting systems, self-erectable structures, clamps, thermostatic devices, and biomedical applications (5,21–23).

Andreasen suggested the first clinical application of Ni–Ti SMA in 1971. He suggested that nitinol wire was useful for orthodontics by reason of its superelasticity and good corrosion resistance (24). Since then, Ni–Ti alloys have been used in a broad and continually expanding array of biomedical applications, including various prostheses and disposables used in vascular and orthopedic surgery. Medical interventions have themselves been driven toward minimally invasive procedures by the creation of new medical devices, such as guide wires, cardiovascular stents, filters, embolic coils, and endoscopic surgery devices. The Ni–Ti SMA stent was first introduced in 1983 when Dotter (25) and Cragg (26) simultaneously published the results of their experimental studies. However, their studies were unsuccessful because of the unstable introduction system and the intimal hyperplasia in the stent-implanted region (27). In 1990, Rauber et al. renewed the effort to use a Ni–Ti alloy as a stent, significantly reducing intimal hyperplasia by using a transcatheter insertion method (28). In 1992, Josef Rabkin reported successful results in the treatment of obstructions in vascular and nonvascular systems in 268 patients (29). In 1989, Kikuchi reported that a guidewire constructed from kink-resistant titanium–nickel alloy was helpful for angiography and interventional

procedures (30). Guidewires are used for needles, endoscopes, or catheters, to gain access to a desired location within the human body. In 1989, the U.S. Food and Drug Administration approved the use of a Mitek anchor constructed of nitinol for shoulder surgery (31). Since then, many devices and items have been developed with nickel–titanium SMAs.

NICKEL–TITANIUM SHAPE MEMORY ALLOY

Physical Properties

Some of the physical properties of 55-Nitinol are listed in Table 1 (32,33). Nitinol has good impact properties, low density, high fatigue strength, and a nonmagnetic nature. The excellent malleability and ductility of nitinol enable it to be manufactured in the form of wires, ribbons, tubes, sheets, or bars. It is particularly useful for very small devices.

Phase Diagram and Crystal Structures

A Ti–Ni equilibrium phase diagram (34) is very useful for understanding phase transformation and alloy design; a modified one is shown in Fig. 4 (35). There is a triangular region designated "TiNi" near the point of equiatomic composition. The left slope (solubility limit) is nearly vertical with temperature. This means that a precipitation-hardening process cannot be used on the Ti-rich side in bulk alloys. By contrast, the right slope is less steep than the left. Therefore, the precipitation-controlling process can adjust transformation temperatures for practical application of SMAs on the Ni-rich side. The crystal structure of the upper part of this triangle, $> 1090\ ^\circ$C, is body centered cubic (bcc). The lower part is a CsCl-type ordered structure (B2) from 1090 °C to room temperature. A schematic atomic configuration of the B2 structure is shown in Fig. 5 (36). In 1965, Wang determined the lattice constant of the B2 crystal as $a_0 = 3.01$ Å (6). He proposed that the Ni–Ti crystal structure is not a simple CsCl-type structure, but has a disordered 9 Å superlattice and an ordered 3 Å CsCl-type sublattice. As the temperature is lowered, the ordered CsCl structure is slightly tilted instantaneously and cooperatively into a close-packed structure, called martensite, with a 2D dimensional close-packed plane (basal plane) (6,37). The martensite unit cell is described as a monoclinic (B19′) configuration, as shown in Fig. 6.

Table 1. Some of the Physical and Mechanical Properties of Nominal 55-Nitinol[a]

Density	6.45 g/cm^3
Melting point	1310 °C
Magnetic permeability coefficient	< 1.002
Electrical resistivity	
20 °C	80 μΩ · cm
900 °C	132 μΩ · cm
Thermal expansion	10.4×10^{-6}/ °C
Hardness,	
950 °C furnace cooled	89 R_B
950 °C quenched	89 R_B
Yield strength	103–138 MPa (15–20 $\times$ 10^3 psi)
U.T.S.	860 MPa (125 $\times$ 10^3 psi)
Elongation	60%
Young's modulus	70 GPa (10.2 $\times$ 10^6 psi)
Shear modulus	24.8 GPa (3.6 $\times$ 10^6 psi)
Poisson's ratio	0.33
Fatigue (Moore test) stress 10^7 counts	480 MPa (70 $\times$ 10^3 psi)
Charpy impact	
Unnotched (RT)[b]	155 ftlb
Unnotched (−80 °C)	160 ftlb
Notched (RT)	24 ftlb
Notched (−80 °C)	17 ftlb

[a]Reproduced with permission from Biocompatibility of Clinical Implant Materials volume I, Ed. By D. F. Williams, 1981, Table 2 on page 136, Castleman L. S. and Motzkin S. M., copyright CRC press, Boca Raton Florida. See Refs. (32) and (33).
[b]Room temperature = RT.

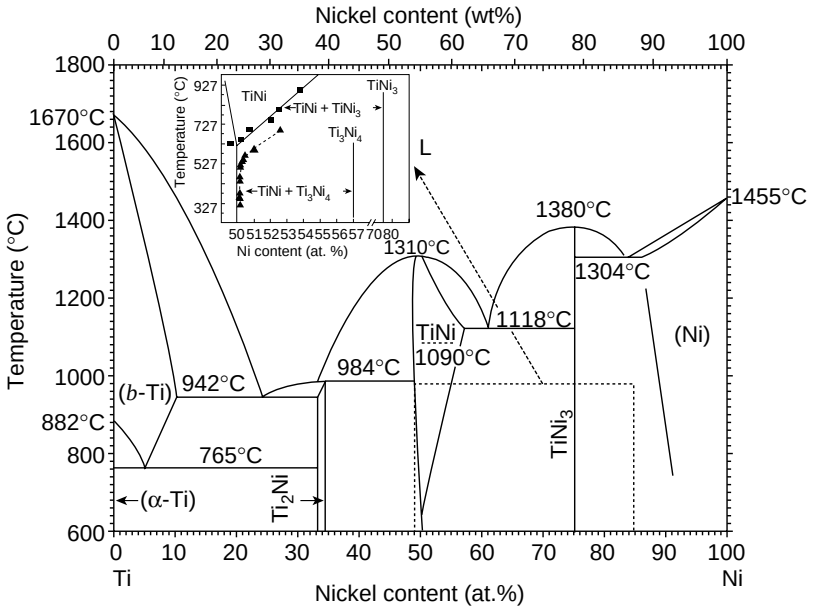


Figure 4. Phase diagram of a Ti–Ni alloy and details of the TiNi and $TiNi_3$ phases (35). (Reproduced with permission from Binary Alloy Phase Diagrams, 2nd ed., Vol. 3, 1990, Phase diagram of a Ti-Ni alloy on page 2874, T. B. Massalski, H. Okamoto, P. R. Subramanian, and L. Kacprzak, ASM International.)

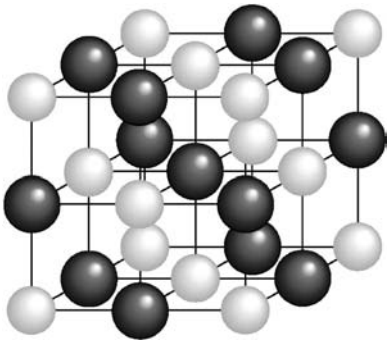


Figure 5. Schematic 3D diagram of the Ni–Ti atomic model in the stable high temperature phase (CsCl-type structure; lattice constant; $a_0 = 3.01$ Å).

The twin-type stacking of the thermally induced martensite structure shown on the left (a) has a readily deformable crystalline arrangement, from the twin structure to the detwinned structure shown on the right (b) (9,38). Diagram (b) of the detwinned structure shows relatively planar atomic stacking layer by layer alternately along the {111} basal plane of the deformed martensite crystal (39). Since martensitic transformation in Ni–Ti SMAs demonstrates an abnormal heat capacity change, it is regarded as a crystallographic distortion instead of a crystallographic transformation. The Ni–Ti martensite transformation is accompanied by a large latent heat of enthalpy ($\Delta H \sim 4{,}150$ J/mol). This extraordinarily latent heat of transformation was considered to be owing to a portion of the electrons undergoing a "covalent-to-metallic" electron-state transformation (11).

Thermomechanical Properties

The mechanical properties of Ni–Ti SMAs are closely dependent on the testing temperature. If a mechanical stress is applied to the SMA below the TTR, then the metastable parent structure of the Ni–Ti alloy is susceptible to transformation into the martensite. However, if the testing temperature exceeds the TTR, then, in the absence of stress, the reverse transformation happens. Figure 7 shows an example of a uniaxial compressive stress-strain curve of a Ni–Ti alloy above its TTR, which shows its superelasticity (40).

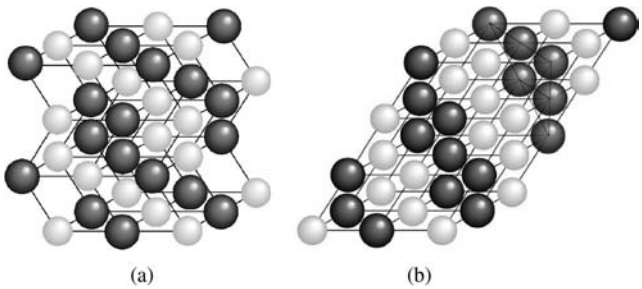


Figure 6. Schematic 3D diagram of the Ni–Ti atomic stacking model of low temperature stable monoclinic structured martensite (a) twin-type stacking of martensite, (b) detwinned-type stacking of martensite).

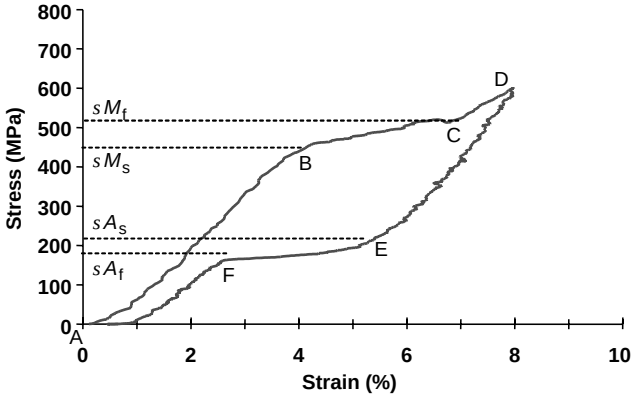


Figure 7. Compressive stress–strain curve of a heat-treated 6-mm-diameter Ni–Ti rod at 4 °C. Three distinct stages are observed on the stress–strain curve (σM_f: stress-induced martensite finishing stress, σM_s: stress-induced martensite starting stress, σA_s: parent phase starting stress, σA_f: parent phase finishing stress) (40).

With stress below the martensite starting stress (σM_s), the Ni–Ti alloy behaves in a purely elastic way, as shown in section AB. As soon as the critical stress is reached at point B, corresponding to stress level σM_s, forward transformation (parent phase-to-martensite) is initiated and SIM starts to form. The slope of section BC (upper plateau) reflects the ease with which the transformation proceeds to completion, generating large transformational strains. When the applied stress reaches the value of the martensite finishing stress (σM_f), the forward transformation is completed and the SMA is fully in the SIM phase. For further loading above σM_f, the elastic behavior of martensite is observed again until plastic deformation occurs, as represented in section CD. For stress beyond D, the material deforms plastically until fracture occurs. However, if the stress is released before reaching point D, the strain is recovered in several stages. The first stage is elastic unloading of the martensite, as shown in section DE. On arriving at stress σA_s, at E, the reverse martensite transformation starts and the fraction of martensite decreases until the parent phase is completely restored at F. Section FA represents the elastic unloading of the parent phase. If some irreversible deformation has taken place during either loading or unloading, the total strain may not be recovered completely. Owing to the stress differences between σM_f and σA_s and between σM_s and σA_f, a hysteresis loop is obtained in the loading–unloading stress–strain curve. Increasing the test temperature results in an increase in the values of the critical transformation stresses, while the general shape of the hysteresis loop remains the same. The area enclosed by the loading and unloading curves represents the energy dissipated during a stress cycle. As part of the hysteresis loop, both the loading and unloading curves show plateaus, at which point large strains are accommodated on loading, or recovered on unloading, with only a small change in stress (19). This behavior of Ni–Ti SMAs is much like that of natural tissues, such as hair and bone, and results in a "superelastic" ability to withstand and recover from large deforming stresses.

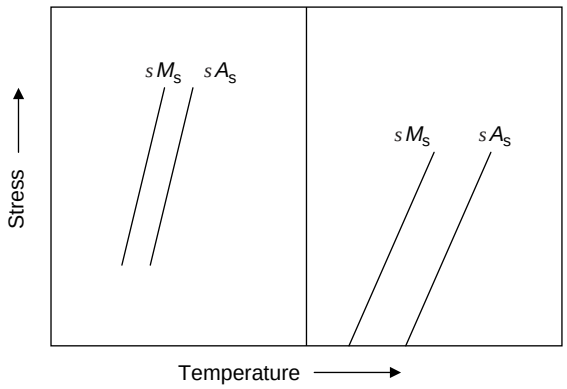


Figure 8. The effect of compressive (a) and tensile (b) loading on martensite formation and disappearance in 20.7% Tl–In alloy (19). (From J. Mat. Sci. Vol. 9, 1974, Figure 2 on page 1537, Krishnan R. V., Delaey L., Tas H. and Warlimont H., Kluwer Academic Publisher. Reproduced with kind permission of Springer Science and Business Media.)

In 1974, Krishnan argued that Burkart and Read had found the effects of compressive and tensile stress on martensite formation and disappearance in Tl–In SMAs (19). The transformation stresses σM_s and σA_s have a linear relation with testing temperature, as shown in Fig. 8 (41). They inferred that σM_s is a linear function of temperature, and the stresses σM_s and σA_s increase with temperature.

Another important thermomechanical property of SMAs is the relationship between the plateau stress of the martensite phase transformation and the enthalpy change of that reaction. As the stress-induced martensitic transformation is a second-order transformation, the amount of transformation depends on its temperature, so the high temperature state has a larger energy barrier of SIM and needs more energy to overcome this larger reverse martensitic transformation barrier. The enthalpy change of the parent phase to martensitic transformation ($\Delta H^{\text{p-m}}$) can be calculated theoretically using the modified Clausius–Clapeyron equation (20), shown in Eq. 1.

$$\frac{\delta\sigma^{\text{p}-\text{m}}}{\delta T} = \frac{\rho\Delta H^{\text{p-m}}}{\varepsilon^{\text{p}-\text{m}} T_{\text{o}}} \tag{1}$$

Where $\Delta H^{\text{p}-\text{m}}$ is the enthalpy change of the parent phase to the martensite phase at T_{o}; $\sigma^{\text{p}-\text{m}}$ is the stress at which stress-induced martensite is formed at the testing temperature, T; ρ is the density of the SMA; and $\varepsilon^{\text{p}-\text{m}}$ is the strain corresponding to complete transformation. $\delta\sigma^{\text{p}-\text{m}}$ and $\varepsilon^{\text{p}-\text{m}}$ can be taken from the stress–strain curves. Kim compared the theoretically calculated $\Delta H^{\text{p}-\text{m}}$ of a Ni–Ti alloy using stress–strain curves and Eq. 1 with an experimentally acquired value (9). He reported that the theoretical value of $\Delta H^{\text{p}-\text{m}}$ for a Ni–Ti alloy calculated from the stress–strain curves was 6.24 cal/g. The experimental value of the enthalpy change ($\Delta H^{\text{p}-\text{m}}$) of an 8% prestrained Ni–Ti wire sample from DSC measurement was 6.79 cal·g^{-1}. Based on this result, he inferred that Ni–Ti alloys undergo thermomechanical-phase transformation by exchanging thermal energy into mechanical energy and vice versa (9).

MANUFACTURING METHODS

Alloy Refining

The Ni–Ti SMAs can be refined using either the vacuum-induction melting method or the consumable arc melting technique. In vacuum-induction melting, a prerequisite in working with Ni–Ti is a high purity graphite crucible. To prevent impurities, the crucible should be connected to the pouring lip mechanically to keep the molten Ni–Ti compound from contacting anything, but the high density, low porosity graphite. Elemental carbon is very reactive with Ni or Ti alone and any contact with either will ruin the purity of the desired sample. However, there is very little reaction with the crucible in the consumable arc melting process. This method yields a product that is relatively free of impurities. Once the Ni–Ti alloy is cast using the melting technique, it is ready for hot or cold working into more practical forms and consecutive annealing treatment (42).

Mechanical Processing

When hot working a piece of Ni–Ti alloy, the temperature should be below that where incipient melting of the secondary phase can occur. This temperature should also be held constant for a period of time sufficient for certain nonequilibrium phases to return to solution, which makes the remaining alloy homogeneous. Andreasen suggested that the optimum hot working temperature is 700–800 °C for forging, extrusion, swaging, or rolling. If cold rolling is desired, then the alloy should be annealed before the oxide is removed (42). The most common form of Ni–Ti alloy is a wire. To make a wire, the Ni–Ti alloy ingot must be rolled into a bar at high temperature. Swaging the bar, followed by drawing, and a final annealing, reduce the alloy to wire form. To soften the wire, it should be annealed between 600 and 800 °C for a short period. When the Ni–Ti alloy is drawn down to 0.8 mm through a carbide die, the maximum reduction in area with each pass should be within 10%. Once this diameter is reached, a diamond die is used to draw the alloy with a 20% area reduction per die. The Ni–Ti alloy is annealed again at 700 °C and allowed to cool to room temperature between passes (42). By contrast, the extrusion method is used for the tube-making process, which enables a substantially greater reduction in cross-sectional area as compared to drawing wire. Laser cutting of Ni–Ti tubes has been used to make vascular stents (43). Most Ni–Ti alloys require a surface finishing procedure after the final machining process, such as chemical leaching, cleaning, rinsing, and surface modification.

Shape Memory Programming

There are two steps in the shape memory programming of a Ni–Ti alloy. First, the Ni–Ti alloy sample must be deformed to the desired shape and put into a constraining mold or fixture. The next step is shape memory heat treatment in a furnace at 400–600 °C. The shape recovery efficiency of a Ni–Ti alloy can be controlled by changing the heat treatment conditions or the degree of deformation. In general, there are three different ways to control the TTR of a Ni–Ti SMA: altering the chemical composition,

changing the heat treatment conditions, and varying the degree of deformation (13).

Chemical Composition Effect on TTR. The shape memory characteristic is limited to Ni–Ti alloys with near-equiatomic composition, as shown in Fig. 4. A pure stoichiometric (50 at%) Ni–Ti alloy will have a nickel content of $\sim$ 55 wt%. Increasing the nickel concentration lowers the characteristic transformation temperature of the alloy. The limit of the nickel concentration for a SMA is $\sim$ 56.5 wt%, owing to the formation of a detrimental second phase. In addition, the shape memory properties of a Ni–Ti alloy can be readily modified by adding ternary elements that are chemically similar to Ni or Ti. Adding a small amount of a transition metal such as Co, Fe, or Cr, instead of Ni, depresses the TTR, such that the SME occurs at well-below ambient temperature (44). When larger ions are substituted for smaller ions, the transformation temperature increases. Concerning ternary additions to alloys, Murakami et al. (45) proposed that the stability of the parent phase is controlled by ion–core repulsive interactions such that when larger ions are substituted for smaller ions, the transformation temperature increases. Based on this hypothesis, substitutions of Au and Zr should increase the recovery temperature of Ni–Ti alloys, Al and Mn should decrease it, and Co and Fe should cause little change. The effects of Au, Zr, Al, and Mn were predicted correctly, but those of Co and Fe were not. Similarly, Morberly suggested that if $>$ 7.5% copper is added to a Ni–Ti alloy, up to 30%, the addition of Cu increases and narrows the TTR (46).

Mechanical Deformation Effect on the TTR. Many investigators have reviewed the effect of mechanical deformation on the TTR (5,36,47). They found that the degree of deformation affects the TTR of a SMA, and the stress slope ($d\sigma/dT$) is a very important fundamental descriptor of SMAs. The residual stress from prior cold work can have a major effect on the transformation behavior. As a result, retention of the parent phase is a function of the stress and heat treatment history. Lee et al. reported that bending beyond the yielding point broadened the TTR and increased the stored internal energy (48). Figure 9 shows an example of transition temperature variation with respect to uniaxial prestrain of a Ni–Ti alloy wire (13). When the prestrain $>$ 8%, the shape recovery transition temperature (A_s) and the martensite starting temperature (M_s) are increased with increasing prestrain. However, the enthalpy change of the cooling cycle is almost the same because most stored internal energy in SIM is already liberated during the heating cycle (36).

Heat Treatment Effect on the TTR. The TTR of a Ni–Ti alloy can be controlled by the final annealing temperature and time. Kim insisted that a higher annealing temperature gives a lower transition temperature and a wider TTR (9). Moreover, he showed that a larger grain size has a lower transition temperature because the annealed large grains have much more transformable volume than smaller grains, so they need more energy for second-phase nucleation and growth inside the grain (49).

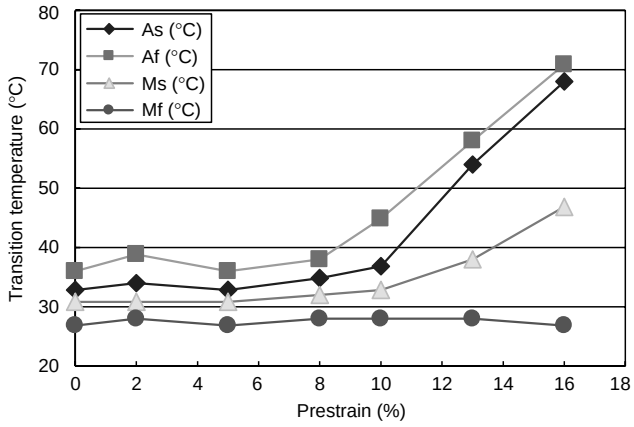


Figure 9. Transition temperatures of prestrained Ni–Ti alloy wire (13).

Figure 10 shows an example of the heat treatment temperature effect on SME (40). When a Ni–Ti rod is heat treated for 30 min at 600 °C, the rod shows superelasticity at room temperature. This indicates that the TTR is lower than the testing temperature. By contrast, when a Ni–Ti rod is heat treated for 30 min at $<$ 500 °C, the rod shows SME at room temperature, which suggests that the TTR is higher than the testing temperature. These results clearly show that the SME is closely related to the heat treatment temperature (9,50).

Methods of Measuring Transition Temperatures

There are many measurable parameters that accompany the shape memory transformation of a Ni–Ti alloy, for example, hardness, velocity of sound, damping characteristic, elastic modulus, thermal expansion, electrical resistivity, specific heat, latent heat of transformation, thermal conductivity, and lattice spacing. Of these, the electrical resistivity and latent heat of transformation are useful for measuring the TTR of a SMA.

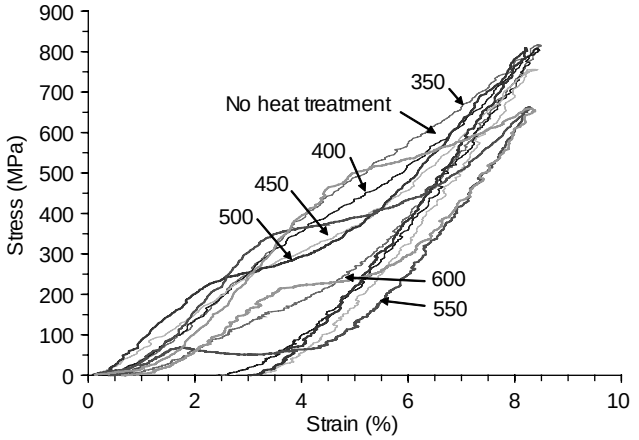


Figure 10. The room temperature compression stress–strain curves of heat-treated ϕ6-mm Ni–Ti rods for 30 min at 350–600 °C. The numbers pointing to the graphs are the annealing temperatures (40).

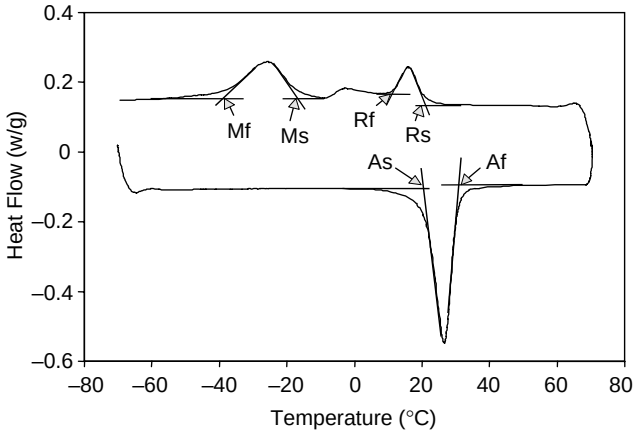


Figure 11. A cyclic DSC curve of the specific heat versus temperature for a Ni–Ti alloy wire from −70 to 70 °C. The lower and upper parts of the cyclic curve represent heating and cooling processes, respectively. (A_s: shape recovery starting temperature, A_f: shape recovery finishing temperature, M_s: martensitic transformation starting temperature, M_f: martformation starting temperature, R_f: R phase transformation finishing temperature) (40).

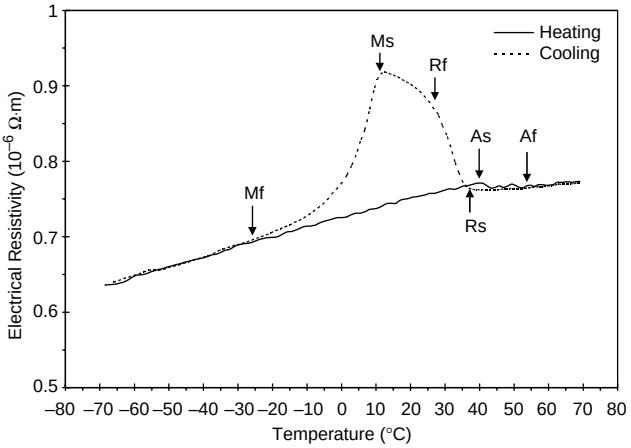


Figure 12. The electrical resistivity versus temperature curve of a ϕ1.89 mm Ni–Ti alloy wire that was heat treated at 550 °C (53).

DSC Measurement. Differential scanning calorimetry (DSC) is a thermal analysis technique that determines the specific heat, heat of fusion, heat of reaction, or heat of polymerization of materials. It is accomplished by heating or cooling a sample and reference under such conditions that they are always maintained at the same temperature. The additional heat required by the sample to maintain it at the same temperature is a function of the observed chemical or physical change (50). Figure 11 shows a typical DSC curve of the specific heat change of a Ni–Ti alloy (40). The lower curve is the heating curve and the upper one is the cooling curve. Each peak represents a phase transformation during the thermal cycle. The area under the curve represents the enthalpy change (ΔH) during the phase transformation. The arrows on Fig. 11 indicate the transition temperatures. The advantage of DSC measurement is that samples can be small and require minimal preparation. In addition, it can detect the residual strain energy, diffusing DSC peaks (51).

Electrical Resistivity Measurement. The shape memory transition temperature can also be determined from the curve of the electrical resistance versus temperature using a standard four-probe potentiometer within a thermal scanning chamber. In 1968, Wang reported the characteristic correlation between the shape memory phase transformations of a Ni–Ti alloy and the irreversible electrical resistivity curves (52). He proposed that the electrical resistivity curve in the same temperature range has a two-step process on cooling, that is, from the parent phase via R-phase to the final martensite phase, and a one-step process on heating, that is, from the martensite to the parent phase. Figure 12 plots the electrical resistivity versus temperature curves of a ϕ1.89 mm Ni–Ti alloy wire that was heat treated at 550 °C for 30 min. During the heating process, the electrical resistivity increases up to temperature A_s, and then it decreases until temperature A_f is reached. This suggests the restoration of the parent structure accompanying this resistivity change. During the cooling cycle, however, a triangular curve appears. The increasing part of this triangular curve from R_s to R_f represents the formation of an intermediate R phase resulting in a further increase in electrical resistivity. The decreasing part represents the thermal energy absorption of the martensitic phase transformation.

Corrosion Resistance

The Ni–Ti SMAs are an alloy of nickel, which is not corrosion resistant in saline solutions, such as seawater, and titanium, which has excellent corrosion resistance under the same conditions. The corrosion resistance of Ni–Ti alloys more closely resembles that of titanium than that of nickel. The corrosion resistance of Ni–Ti alloys is based mainly on the formation of a protective oxide layer, which is called passivation (9). If the alloy corrodes, then breakdown of the protective oxide film on the alloy's surface occurs locally or generally, and substantial quantities of metallic ions are released into the surrounding solution. Therefore, corrosion resistance is an important determinant of biocompatibility (54–56). The Pourbaix diagram is a useful means of measuring corrosion. It is a potential versus pH diagram of the redox and acid–base chemistry of an element in water. It is divided into regions where different forms of the metal predominate. The three regions of interest for conservation are corrosion, immunity, and passivity. The diagram may indicate the likelihood of passivation (or corrosion) behavior of a metallic implant *in vivo*, as the pH varies from 7.35 to 7.45 in normal extracellular fluid, but can reach as low as 3.5 around a wound site (57). An immersion test is also used for determining the concentration of released metallic ions, corrosion rates, corrosion types, and passive film thickness in saline, artificial saliva, Hank's solution, physiological fluids, and so on (58).

Some surface modifications have been introduced to improve the corrosion properties of Ni–Ti alloys, and prevent the dissolution of nickel. These include titanium

nitride coating of the Ni–Ti surface and chemical modification with coupling agents for improving corrosion resistance. However, when the coating on a Ni–Ti alloy is damaged, corrosion appears to increase in comparison with an uncoated alloy (56). Laser surface treatment of Ni–Ti leads to increases in the superficial titanium concentration and thickness of the oxide layer, improving its cytocompatibility up to the level of pure titanium (9). Electropolishing methods and nitric acid passivation techniques can improve the corrosion resistance of Ni–Ti alloys owing to the increased uniformity of the oxide layer (59).

Biocompatibility

Biocompatibility is the ability of a material or device to remain biologically inactive during the implantation period. The purpose of a biocompatibility test is to determine potential toxicity resulting from contact of the device with the body. The device materials should not produce adverse local or systemic effects, be carcinogenic, or produce adverse reproductive or developmental effects, neither directly nor through the release of their material constituents (60). Therefore, medical devices must be tested for cytotoxicity, toxicity, specific target-organ toxicity, irritation of the skin and mucosal surfaces, sensitization, hemocompatibility, short-term implantation effects, genotoxicity, carcinogenicity, and effects on reproduction.

The biocompatibility of a Ni–Ti alloy must include the biocompatibility of the alloy's constituents. As Ni–Ti alloys corrode, metallic ions are released into the adjacent tissues or fluids by some mechanisms other than corrosion (61). Although Ni–Ti alloys contain more nickel than 316L stainless steel, Ni–Ti alloys show good biocompatibility and high corrosion resistance because of the naturally formed homogeneous TiO_2 coating layer, which has a very low concentration of nickel. Although Ni–Ti alloys have the corrosion resistance of titanium, the passivated oxide film will dissolve at some rate; furthermore, the oxide layer does not provide a completely impervious barrier to the diffusion of nickel and titanium ions (62,63).

Many investigators have reported on the biocompatibility of Ni–Ti alloys. Comparing the corrosion resistance of common biomaterials, the biocompatibility of Ni–Ti ranks between that of 316L stainless steel and Ti6A14V, even after sterilization. Some of these findings are listed here. Thierry found that electropolished Ni–Ti and 316L stainless steel alloys released similar amounts of nickel after a few days of immersion in Hank's solution (64). Trepanier reported that electropolishing improved the corrosion resistance of Ni–Ti stents because of the formation of a new homogeneous oxide layer (59). In a short-term biological safety study, Wever found that a Ni–Ti alloy had no cytotoxic, allergic, or genotoxic activity and was similar to the clinical reference control material AISI 316 LVM stainless steel (65). Motzkin showed that the biocompatibility of nitinol is well within the limits of acceptability in tissue culture studies using human fibroblasts and buffered fetal rat calvaria tissue (66). Ryhanen reported that nitinol is nontoxic, nonirritating, and very similar to stainless steel and Ti–6Al–4V alloy in an *in vivo* soft tissue and inflammatory response study (67). Castleman found no significant histological compatibility differences between nitinol and Vitallium (Co–Cr alloy) (68). However, Shih reported that nitinol wire was toxic to primary cultured rat aortic smooth muscle cells in his cytotoxicity study using a supernatant and precipitate of the corrosion products (69). Moreover, he found that the corrosion products altered cell morphology, induced cell necrosis, and decreased cell numbers.

MEDICAL DEVICES

The Ni–Ti alloys have been used successfully for medical and dental devices because of their unique properties, such as SME, superelasticity, excellent mechanical flexibility, kink resistance, constancy of stress, good elastic deployment, thermal deployment, good corrosion resistance, and biocompatibility. Recently, Ni–Ti alloys have found use in specific devices that have complex and unusual functions, for example, self-locking, self-expanding, or compressing implants that are activated at body temperature (58). Some popular examples of Ni–Ti medical devices have been selected and are reviewed below.

Orthodontic Arch Wires

A commercially available medical application of nitinol is the orthodontic dental arch wire for straightening malpositioned teeth, marketed by Unitek Corporation under the name Nitinol Active-Arch (70). This type of arch wire, which is attached to bands on the teeth, is intended to replace the traditional stainless steel arch wire. Although efforts have been made to use the SME in orthodontic wires (71), the working principle of Nitinol Active-Arch wire is neither the SME nor pseudoelasticity, but the rubber-like behavior and relatively low Young's modulus (30 GPa) of nitinol in the martensitic condition. This modulus is very low in comparison with the modulus of stainless steel (200 GPa). Comparing the bending moment change of nitinol and stainless steel wire undergoing a constant change in deflection (72), stainless steel wire shows a much larger change in moment than the moment change of nitinol wire. Clinically, this means that for any given malocclusion nitinol wire will produce a lower, more constant force on the teeth than would a stainless steel wire of equivalent size. Figure 13 shows a clinical example of orthodontic treatment using a superelastic Ni–Ti arch wire (73). This wire showed faster movement of teeth and shorter chair time than conventional stainless steel wire.

Guidewires

One typical application of superelasticity is the guidewires that are used as guides for the safe introduction of various therapeutic and diagnostic devices. A guidewire is a long, thin metallic wire that is inserted into the body through a natural opening or a small incision. The advantages of using superelastic guide wire are the improvement in kink resistance and steerability. A kink in a guidewire creates a difficult situation when the time comes to remove it from a

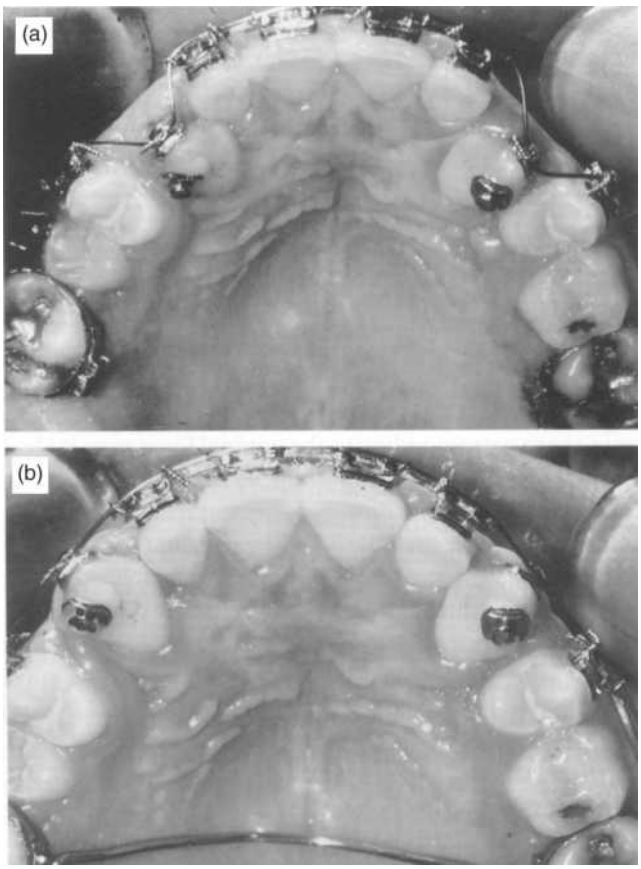


Figure 13. Orthodontic treatment using a Ni–Ti superelastic arch wire. (a) Malaligned teeth before treatment and (b) normally aligned teeth after the first stage of treatment (73). (Reprinted with permission from Shape memory materials, Ed. By K. Otsuka and C. M. Wayman, 1998, Figure 12.3 on page 270, S. Miyazaki, Cambridge University Press.)

complex vascular structure. The enhanced twist resistance and flexibility make it easier for the guidewire to pass to the desired location (74). Figure 14 shows the tip of a guidewire. The curved "J" tip of the guidewire makes it easy to select the desired blood vessel.

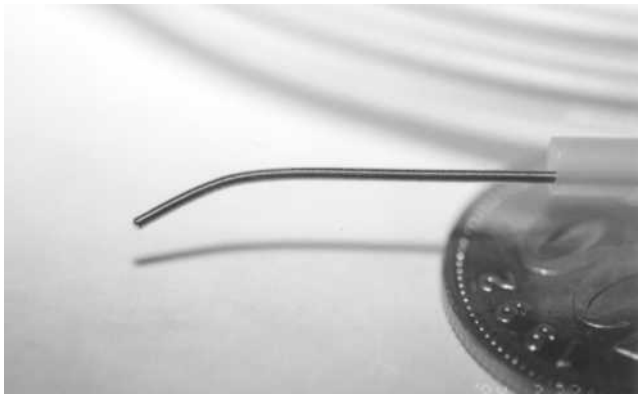


Figure 14. Photograph of the tip of a commercial Ni–Ti guidewire (FlexMedics, USA) (75).

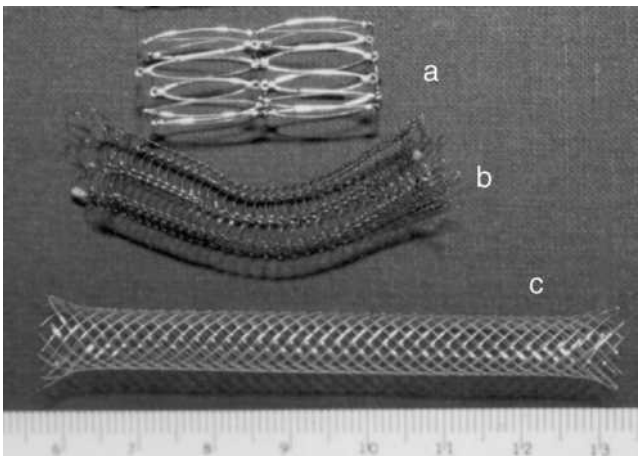


Figure 15. Commercial Ni–Ti stents (a) Gianturco stent, (b) self-expanding nitinol stent with the Strecker stent design, (c) Wall stent (76).

Stents

A stent is a slender metal mesh tube that is inserted inside a luminal cavity to hold it open during and after surgical anastomosis. Superelastic nitinol stents are very useful for providing sufficient crush resistance and restoring lumen shape after deployment through a small catheter (25–27). Figure 15 shows three examples of commercial self-expandable Ni–Ti superelastic stents: a Gianturco stent for the venous system, a Strecker stent for a dialysis shunt, and a Wall stent for a hepatic vein. Figure 16 shows the moment of expansion of a Ni–Ti self-expandable stent being deployed from the introducer. The driving force of the self-expanding stent is provided by the superelasticity of the Ni–Ti alloy. Some clinical limitations of Ni–Ti stents remain unresolved and require further development; these are the problems of intimal hyperplasic and restenosis (78).

Orthopedic Applications

Dynamic compression bone plates exhibiting the SME are one of the most popular orthopedic applications of nitinol, followed by intramedullary fixation nails. Fracture healing

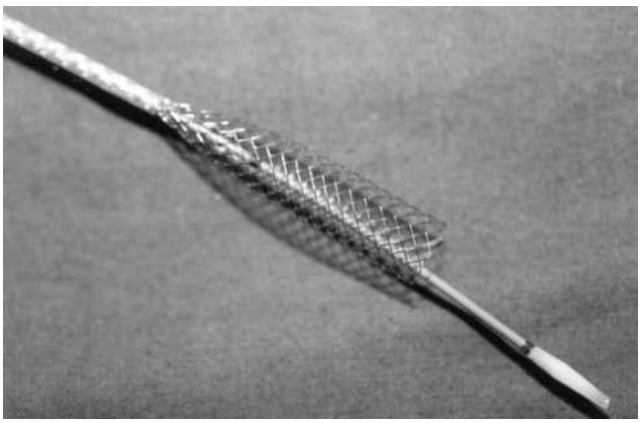


Figure 16. Deployment of a commercial Ni–Ti self-expandable stent (Taewoong Medical, Korea) (77).

in long bones can be accelerated when bone ends are held in position with compression between the bone fragments. Using this method, the undesirable surface damage and wear of the holes that occur in a conventional dynamic bone plate are avoided, while continuous compression is assured, even if bone resorption occurs at the fracture sites. The effect continues as long as the original shape is not reached (79).

Historically, the first orthopedic application of a SMA was a Nitinol Harrington instrument for scoliosis treatment that was introduced in 1976 by Schmerling (80), which enabled the surgeon to restore any relaxed corrective force postoperatively simply by the external application of heat. In addition, it could be used initially to apply a more appropriate set of corrective forces. Figure 17 shows an example of a Ni–Ti shape memory clamp in small bone surgery (81). Six months after surgery, a non-union was present, although the outcome in this patient was assessed as good.

CONCLUSIONS

A shape memory alloy is a metallic substance that has a memory for shape combined with superelasticity. The mechanisms of a nickel–titanium alloy's shape memory effect and superelasticity are described based on thermally induced or stress induced martensite phase transformations. Some of the physical properties of nickel–titanium alloys and a phase diagram are included for reference. The thermomechanical characteristics, corrosion properties, and biocompatibility of Ni–Ti shape memory alloys are reviewed for the design of shape memory devices. Manufacturing methods, including refining, processing, shape memory programming, and transformation temperature range measuring methods are summarized for practical applications. Finally, some applications in medical devices are reviewed as examples of current trends in the use of shape memory alloys. In conclusion, Ni–Ti shape memory alloys are a very useful biocompatible material because of their unique mechanical properties and good corrosion resistance. A better understanding of shape memory alloys should allow further developments in this area.

Figure 17. Failed arthrodesis of the carpometacarpal joint when only one titanium–nickel (TiNi) clamp was used (81). (From Arch. Orthop. Trauma Surg., Vol. 117, 1998. Figure 1 on page 342, Musialek J., Filip P. and Nieslanik J. Reproduced with kind permission of Springer Science and Business Media.)

BIBLIOGRAPHY

Cited References

1. Properties and selection. Metal handbook 8th edition volume 1. American Society for Metals; 1961. p 1.
2. Jena AK, Chaturvedi MC. Phase transformation in materials. Prentice Hall; 1992. p 1–3.
3. Park JB, Kim YK. Metallic biomaterials. In: Bronzino JD, editor. The biomedical engineering handbook. 2nd ed. Volume 1, CRC Press; 2000. p 37-1–37-20.
4. Buehler WJ, Gilfrich JV, Wiley RC. Effect of low-temperature phase changes on the mechanical properties of alloys near composition TiNi. J Appl Phys 1963;34:1475–1477.
5. Wayman CM, Duerig TW. An introduction to martensite and shape memory. In: Duerig TW, Melton KN, Stoeckel D, Wayman CM, editors. Engineering aspects of shape memory alloys. Butterworth-Heinemann; 1990. p 3–20.
6. Wang FE, Buehler WJ, Pickart SJ. Crystal structure and a unique martensite transition on TiNi. J Appl Phys 1965;36: 3232–3239.
7. Ling CH, Kaplow R. Stress-induced shape changes and shape memory in the R and Martensite transformations in equiatomic NiTi. Metal Trans A 1981;12A:2101–2111.
8. In: Nishiyama Z, Fine ME, Meshii M, Wayman CM, editors. Martensitic Transformation. London: Academic Press; 1978. p 1–13.
9. Kim YK. Thermo-mechanical study of annealed and laser heat treated nickel–titanium alloy dental arch wire. Ph.D. dissertation, University of Iowa, Iowa, Dec. 1989.
10. Wayman CM, Bhadeshia H. Phase transformations, Nondiffusive. In: Cahn RW, Haasen P, editor. Physical Metallurgy. 4th ed. Volume 2, North-Holland: 1996. p 1507–1554.
11. Wang FE, Pickart SJ, Alperin HA. Mechanism of the TiNi transformation and the crystal structures of TiNi-II and TiNi-III phases. J Appl Phys 1972;43:97–112.
12. Otsuka K, Wayman CM, Nakai K, Sakamoto H, Shimizu K. Superelasticity effects and stress-induced martensite transformations in Cu–Al–Ni alloys. Acta Metallurgica 1976;24: 207–226.
13. Kim YK, Doo JK, Park JP. The application of shape memory alloy to abdominoscopic suture needles, In: Shin KS, Yoon JK, Kim SJ, editors. Proceeding of 2nd Pacific RIM International conference on Advanced Materials and Processing. Korean Institue of Metals and Materials; 1995. p 1691–1696.
14. Oelander A. Z Kristallogr 1932;83A:145. as cited in Lieberman DS. Crystal geometry and mechanisms of phase transformations in crystalline solids. In: Aaronson HI, editor. Phase Transformations. American Society for Metals; 1970. p 1–58.
15. Greninger AB, Mooradian VG. Strain transformation in metastable beta copper-zinc and beta copper-tin alloys. Am Inst Mining Met Eng 1937;19:867.
16. Bush RE, Leudeman RT, Gross PM. Alloys of improved properties. AMRA CR 65-02/1, AD629726, U.S. Army Materials Research Agency, 1966.
17. Kurdjumov GV, Khandros LG. Dokl Akad Nauk SSSR 1949;66:211. (as cited in Delaey L, Krishnan RV, Tas H, Warlimont H. Review: thermoelasticity, pseudoelasticity and memory effects associated with martensitic transformations. J Mater Sci 1974;9:1521–1535.
18. Delaey L, Krishnan RV, Tas H, Warlimont H. Review Thermoelasticity, pesudoelasticity and the memory effects associated

with martensitic transformations Part 1 Structural and microstructural changes associated with the transformations. J Mat Sci 1974;9:1521–1535.
19. Krishnan RV, Delaey L, Tas H, Warlimont H. Review Thermoelasticity, pesudoelasticity and the memory effects associated with martensitic transformations Part 2 The macroscopic mechanical behaviour. J Mater Sci 1974;9:1536–1544.
20. Warlimont H, Delaey L, Krishnan RV, Tas H. Review Thermoelasticity, pesudoelasticity and the memory effects associated with martensitic transformations Part 3 Thermodynamics and kinetics. J Mater Sci 1974;9:1545–1555.
21. Melton KN. General applications of SMA's and smart materials. In: Duerig TW, Melton KN, Stoeckel D, Wayman CM, editors. Engineering aspects of shape memory alloys. Butterworth-Heinemann; 1990. p 220–239.
22. Miayazaki S. Medical and dental applications of shape memory alloys, In: Duerig TW, Melton KN, Stoeckel D, Wayman CM, editor. Engineering aspects of shape memory alloys. Butterworth-Heinemann; 1990. p 267–281.
23. Filip P. Titanium-Nickel shape memory alloys in medical applications. In: Brunette DM, Tengvall P, Textor M, Thomsen P, editor. Titanium in Medicine. Springer; 2001. p 53–86.
24. Andreasen GF, Hilleman TB. An evaluation of 55 cobalt substituted Nitinol wire for use in orthodontics. JADA 1971;82: 1373–1375.
25. Dotter CT, Bushmann RW, McKinney MK, Rosch J. Transluminal expandable nitinol coil stent grafting: preliminary report. Radiology 1983;147:259–260.
26. Cragg A, Lund G, Rysavy J, Castaneda F, Castaneda-Zuniga W, Amplatz K. Nonsurgical placement of arterial endoprostheses: a new technique using nitinol wire. Radiology 1983;147: 261–263.
27. Rösch J, Keller FS, Kaufman JA. The Birth, Early Years, and Future of Interventional Radiology. JVIR 2003;14(7):841–853.
28. Rauber K, Franke C, Rau WS, Syed Ali S, Bensmann G. Perorally insertable endotracheal stents made from NiTi memory alloy - an experimental animal study. Rofo Fortschr Geb Rontgenstr Neuen Bildgeb Verfahr 1990;152(6):698–701.
29. Rabkin JE, Germashev V. The Rabkin nitinol coil stent: a five-year experience. In: Castaneda-Zuniga WR, Tadavarthy SM, editors. Interventional Radiology, 2nd ed. Williams & Wilkins; 1992. p 576–581.
30. Kikuchi Y, Graves VB, Strother CM, McDermott JC, Babel SG, Crummy AB. A new guidewire with kink-resistant core and low-friction coating. Cardiovasc Intervent Radiol 1989;12(2): 107–109.
31. Kauffman GB, Mayo I. The story of Nitinol: the serendipitous discovery of the memory metal and its applications. Chem Educator 1997;2(2):S1430–4171; http://chemeducator.org/bibs/0002002/00020111.htm, Feb.2. 2005.
32. Castleman LS, Motzkin SM. The Biocompatibility of Nititnol. In: Williams DF, editor. Biocompatibility of Clinical Implant Materials volume I. CRC Press; 1981. p 129–154.
33. Cross WB, Karitos AH, Wasilewski RJ. Nitinol characterization study. NASA CR-1433, National Aeronautics and Space Administration, Houston. 1969.
34. Buehler WJ, Wiley RC. TiNi-ductile intermetallic compound. Trans ASM 1962;55:269–276.
35. Otsuka K, Kakeshita T. Science and Technology of Shape memory Alloys: New Developments. MRS Bull 2002;27(2): 91–100.
36. Kim YK. The study of the shape recovery temperature change of cold-worked nickel-titanium alloys. Inje J 1994;10(1): 341–352.
37. Aboelfotoh MO, Aboelfotoh HA, Washburn J. Observations of pretransformation lattice instability in near equiatomic NiTi alloy. J Appl Phys 49(10): 1978; 5230–5232.
38. Ling HC, Kaplow R. Phase transitions and shape memory in NiTi. Metal Trans A 1980;11A:77–83.
39. Chandra K, Purdy GR. Obseravation of thin crystals of TiNi in premartensite states. J Appl Phys 19(5): 1968; 2176–2181.
40. Shin SH. The study of heat-treatment temperature effect on hardness and compressional properties of nickel-titanium alloy. Master. dissertation, Inje University, Korea, Dec. 1998.
41. Burkart MW, Read TA. Trans Met Soc AIME 1953;197:1516. (as cited in Krishnan RV, Delaey L, Tas H, Warlimont H. Review Thermoelasticity, pesudoelasticity and the memory effects associated with martensitic transformations Part 2 The macroscopic mechanical behaviour. J Mat Sci 1974;9:1536–1544.
42. Andreasen GF, Fahl JL. Alloys, Shape Memory. In: Webster JG, editor. Encyclopedia of Medical Devices and Instrumentation. Volume 1, New York: Wiley-Interscience; 1988. p 15–20.
43. Thierry B, Merhi Y, Bilodeau L, Trepanier C, Tabrizian M. Nitinol versus stainless steel stents: acute thrombogenicity study in an ex vivo porcine model. Biomaterials 2002;23:2997–3005.
44. Moberly WJ, Melton KN. Ni–Ti–Cu shape memory alloys. Engineering Aspects of Shape Memory Alloys. London: Butterworth-Heinemann; 1990. p 46–57.
45. Murakami Y, Asano N, Nakanishi N, Kachi S. Phase relation and kinetics of the transformations in Au–Cu–Zn thernary alloys. Jpn J Appl Phys 1967;6:1265–1271.
46. Gil FJ, Planell JA. Effect of copper addition on the superelastic behavior of Ni–Ti shape memory alloys for orthodontic applications. J Biomed Mater Res Appl Biomat 1999;48:682–688.
47. Goldstein D, Kabacoff L, Tydings J. Stress effects on Nitinol phase transformations. J Metals 1987;39(3):19–26.
48. Lee JH, Park JB, Andreasen GF, Lakes RS. Thermo mechanical study of Ni–Ti alloys. J Biomed Mater Res 1988;22: 573–588.
49. Kim YK. The Grain size distribution study of heat treated Ni–Ti alloy. Inje J 1993;9(2):857–868.
50. Differential Scanning Calorimetry, Dept. of Polymer Science, University of Southern Mississippi. Available at http://www.psrc.usm.edu/macrog/dsc.htm. Accessed Feb. 8. 2005.
51. Harrison JD. Measurable changes concomitant with the shape memory effect transformation, In: Duerig TW, Melton KN, Stoeckel D, Wayman CM, editors. Engineering aspects of shape memory alloys. Butterworth-Heinemann; 1990. p 106–111.
52. Wang FE, DeSavage BF, Buehler WJ, Hosler WR. The irreversible critical range in the TiNi transition. J Appl Phys 1968;39(5):2166–2175.
53. Kim YK. unpublished experimental data (ykkimbme.inje.ac.kr).
54. Cutright DE, Bhaskar SN, Perez B, Johnson RM, Cowan GS, Jr. Tissue reaction to nitinol wire alloy. Oral Surg 1973;35(4): 578–584.
55. Castleman LS, Motzkin SM, Alicandri FP, Bonawit VL. Biocompatibility of Nitinol alloy as an implant material. J Biomed Mater Res 1976;10:695–731.
56. Castleman LS, Motzkin SM. The biocompatibility of Nitinol. In: Williams DF, editor. Biomcompatibility of Clinical Implant Materials. Volume 1, CRC Press; 1981. p 129–154.
57. Park JB. Metallic implant materials. Biomaterials Science and Engineering. Joon Bu Park: Plenum Press; 1984. p 193–233.
58. Ryhaenen J. Biocompatibility Evaluation of Nickel-Titanium Shape Memory Metal Alloy. Academic Dissertation, University hospital of Oulu, on May 7th, 1999.
59. Trepanier C, Tabrizian M, Yahia LH, Bilodeau L, Piron DL. Effect of modification of oxide layer on NiTi stent corrosion resistance. J Biomed Mater Res (Appl Biomater) 1998;43:433–440.
60. Kammula RG, Morris JM. Considerations for the Biocompatibility Evaluation of Medical Devices. Available at http://www.devicelink.com/mddi/archive/01/05/008.html. Medical Device Link. Accessed Feb. 2. 2005.

61. Austian J. Toxicological evaluation of biomaterials: primary acute toxicity screening program. Artif Organs 1977;1:53–60.
62. Mears DC. The use of dissimilar metals in surgery. J Biomed Mater Res 1975;9:133–148.
63. Williams DF. Future prospects for biomaterials. Biomed Eng 1975;10:206–218.
64. Thierry B, Tabrizian M, Trepanier C, Savadogo O, Yahia LH. Effect of surface treatment and sterilization processes on the corrosion behavior of NiTi shape memory alloy. J Biomed Mater Res 2000;51:685–693.
65. Wever DJ, Veldhuizen AG, Sanders MM, Schakenraad JM, Horn V, Jr. Cytotoxic, allergic and genotoxic activity of a nickel–titanium alloy. Biomaterials 1997;18(16):1115–1120.
66. Motzkin SM, Castleman LS, Szablowski W, Bonawit VL, Alicandri FP, Johnson AA. Evaluation of nitinol compatibility by cell culture. Proc. 4th New England Bioeng Conf. New Haven: Yale University; 1976. p 301.
67. Ryhanen J, Kallioinen M, Tuukkanen J, Junila J, Niemela E, Sandvik P, Serlo W. In vivo biocompatibility evaluation of nickel-titanium shape memory metal alloy: muscle and perineural tissue responses and encapsule membrane thickness. J Biomed Mater Res 1998;41(3):481–488.
68. Castleman LS. Biocompatibility of Nitinol alloy as an implant material. Proceeding of the 5th Annual International Biomaterials Symposium. Clemson (SC): Clemson University; 1973.
69. Shih CC, Lin SJ, Chen YL, Su YY, Lai ST, Wu GJ, Kwok CF, Chung KH. The cytotoxicity of corrosion products of nitinol stent wire on cultured smooth muscle cells. J Biomed Mater Res 2000;52:395–403.
70. Andreasen GF. Method and system for orthodontic moving of teeth. US pat. 4,037,324. 1977.
71. Andreasen GF. A clinical trial of alignment of teeth using a 0.019 inch thermal nitinol wire with a transition temperature range between 31 °C and 45 °C. Am J Orthod 1980;78(5):528–537.
72. Andreasen GF, Morrow RE. Laboratory and clinical analyses of nitinol wire. Am J Orthod 1978;73:142–151.
73. Miyazaki S. Medical and dental applications of shape memory alloys. In: Otsuka K, Wayman CM, editor. Shape memory materials. Cambridge University Press; 1998. p 267–281.
74. Mooney MR, Mooney JF, Pedersen WR, Goldenberg IF, Gobel FL. The Ultra-Select guidewire: a new nitinol guidewire for coronary angioplasty. J Invasive Cardiol 1991;3(5):242–245.
75. Kim YK. unpublished photograph (ykkimbme.inje.ac.kr).
76. Kim YK. unpublished photograph (ykkimbme.inje.ac.kr).
77. Kim YK. unpublished photograph (ykkimbme.inje.ac.kr).
78. Palmaz JC, Kopp DT, Hayashi H, Schatz RA, Hunter G, Tio FO, Garcia O, Alvarado R, Rees C, Thomas SC. Normal and stenotic renal arteries: experimental balloon-expandable intraluminal stenting. Radiology 1987;164(3):705–708.
79. Kousbroek R. Shape memory alloys, In: Ducheyne P, editor. Metal and Ceramic Biomaterials, Volume II: Strength and Surface. CRC Press; Chapt. 3. 1984. p 63–90.
80. Schemerling MA, Wilkov MA, Sanders AE, Woosley JE. Using the shape recovery of Nitinol in the Harrington rod treatment of scoliosis. J Biomed Mater Res 1976;10:879–892.
81. Musialek J, Filip P, Nieslanik J. Titanium-nickel shape memory clamps in small bone surgery. Arch Orthop Trauma Surg 1998;117:341–344.

Reading List

Castleman LS, Motzkin SM. The Biocompatibility of Nititnol. In: Williams DF, editor. Biocompatibility of Clinical Implant Materials volume I. CRC Press; 1981. p 129–154.
Kousbrock R. Shape Memory Alloys, In: Ducheyne P, Hastings GW, editors. Metal and Ceramic Biomaterials Volume II. CRC Press; 1984. p 63–90.
Otsuka K, Wayman CM, editors. Shape Memory Materials. Cambridge University Press; 1998.
Nishiyama Z, Fine ME, Meshii M, Wayman CM, editors. Martensitic Transformation. London: Academic Press; 1978.
Jena AK, Chaturvedi MC. Phase Transformation in Materials. New Jersey: Prentice Hall; 1992.
Duerig TW, Melton KN, Stockel D, Wayman CM. Engineering Aspects of Shape Memory Alloys. London: Butterworth-Heinemann; 1990.
Wayman CM, Bhadeshia HKDH. Phase Transformations, Nondiffusive. Cahn RW, Hassen P, editors. Physical Metallurgy. 4th ed. Amsterdam: North-Holland; 1996. p 1507–1554.
Filip P. Titanium–Nickel Shape Memory Alloys in Medical Applications. In: Brunette DM, Tengvall P, Textor M, Thomsen P, editors. Titanium in Medicine. Berlin: Springer; 2001. p 53–86.
Perkins J, editor. Shape Memory Effects in Alloys. New York: Plenum Press; 1975.

See also HIP JOINTS, ARTIFICIAL; SPINAL IMPLANTS.

AMBULATORY MONITORING

HAIBO WANG
AHMAD ELSHARYDAH
RANDALL CORK
JAMES FRAZIER
Louisiana State University

INTRODUCTION

Due to advances in technology, especially computer sciences, ambulatory monitoring with medical instruments has increasingly become an important tool in the diagnosis of some diseases and medical conditions. Some devices used or in development for current clinical practice are shown in Table 1.

The ideal device for ambulatory monitoring should be consistently sensitive, accurate, lightweight, noninvasive, and easy to use. The Holter monitor is a popular device for ambulatory monitoring. Therefore, this article will start with a discussion of the Holter monitor.

AMBULATORY MONITORING WITH A HOLTER DEVICE

A Holter monitor is a continuous recording of a patient's ecocardiogram (ECG) for 24 h as shown in Fig. 1. It was named in honor of Norman J. Holter for his contribution in creating the world's first ambulatory ECG monitor in 1963 (1). Since it can be worn during the patient's regular daily activities, it helps the physician correlate symptoms of dizziness, palpitations, and syncope with intermittent

Table 1. Current Devices for Ambulatory Monitoring

Devices	Uses
Holter monitoring	Cardiac arrhythmia and ischemia
Ambulatory BP monitoring	Hypertension and hypotension
Ambulatory glucose monitoring	Hyperglycemia and hypoglycemia

Figure 1. Holter monitor.

cardiac arrhythmias. When compared with the ECG, which lasts < 1 min. The Holter monitor is more likely to detect abnormal heart rhythm. It can also help evaluate the patient's ECG during episodes of chest pain due to cardiac ischemia. The common clinical applications for Holter monitor are summarized in Table 2 (2,3).

The basic components of a Holter monitor include at least a portable ECG recorder and a Holter analyzer (scanner). Functional characteristics of both components have improved dramatically since the first Holter monitor was developed > 40 years ago.

Portable ECG Recorder

The recorder is a compact, light-weight device used to record an ambulatory patient's three or more ECG leads, typically for 24 h for dysrhythmia or ischemia detection. There are two types of the recorder available on the market: the classical cassette type (tape) or the newer digital type (flash memory card). The cassette recorder uses magnetic tape to record ECG information. The tape needs to be sent to the physician's office for analysis with a scanner to produce a patient's report. The problems with this type of recorder are its limited memory for ECG recording, its inability to transmit ECG information to the service center digitally, and its difficulty in processing the information with computer software. Therefore, it normally takes days to produce a monitoring report for a patient.

The newer digital recorder has an increased memory compared to the classical cassette type, making it possible to extend the monitoring time beyond 24 h if indicated. More importantly, the recorded signs are digital, which can be transmitted to the service center or on-call physician by digital transmission system via a phone line, email, or wireless technology, and can be processed rapidly with computer software. Therefore, the patient's report will be available for the patient's physician much sooner. If indicated, treatment can be started without delay. Additionally, a patient event button has been incorporated in some of newer recorders to allow correlation of symptoms and activity with ECG changes to obtain more clinical data useful for making a correct clinical diagnosis.

Another new development is the Cardiac Event Monitoring (CEM), which is similar to Holter monitor for recording an ambulatory patient's ECG. The difference between them is that the CEM is an event-triggered device, only recording the patient's ECG when they experiences a detectable symptoms (4). As a result, the CEM makes prolonged monitoring possible even with a limited recorder memory.

Table 2. Clinical Application of the Holter Monitor

1. Evaluation of symptomatic events: dizziness, syncope, heart palpitations, fatigue, chest pain, shortness of breath, episodic diaphoresis.
2. Detection of asymptomatic dysrhythmia: asymptomatic atrial fibrillation.
3. Evaluation of rate, rhythm or ECG interval changes during drug therapy.
4. Evaluation for specific clinical situations: postmyocardial infarction, postcoronary bypass surgery, postpercutaneous transluminal coronary angioplasty, postpacemaker implant, first or second degree heart block, possible pacer malfunction, automatic implanted defibrillator functions.
5. Evaluation of ECG changes during specific activities.

Holter Analyzer (Scanner)

There are different types of Holter analyzer systems currently available. The original system for analyzing the Holter cassette was a scanner with manual observer detection. With manual observer detection, the trained technician watches for audiovisual clues of abnormal beats while playing the tape back at 60–120 times real time. The process is time consuming. It requires a skilled technician who can withstand high boredom and fatigue levels to minimize possible human error rate.

The modern Holter analyzer system has been revolutionized due to the application of computer technologies. It is available in a variety of options, such as auto analysis and complete editing capabilities. Some of the newer systems provide easy-to-use professional features that allow rapid review of recorded information, producing a fast, accurate report.

Future Development

Ambulatory Holter monitoring is a valuable tool in patient care and is becoming more and more popular. Integration of computer technology, digital technology, wireless technology, and nanotechnology may lead to an ideal Holter device, which is minimal in size and weight, user-friendly, noninvasive, sensitive and accurate, wirelessly connected to a physician on-call center, and with automatic data analysis capacity. Newer analysis techniques involving fuzzy logic, neural networks and genetic algorithms will also enhance automatic detection of abnormal ECG. Hopefully, such an ideal ambulatory Holter monitor will be available in the near future.

AMBULATORY BLOOD PRESSURE MONITORING

Introduction

An ambulatory blood pressure monitoring device is a noninvasive instrument used to measure a patient's 24 h ambulatory blood pressure as shown in Fig. 2. The first device was developed by Hinman in 1962 (5). He used a

Figure 2. Ambulatory blood pressure monitoring device.

microphone placed over the brachial artery distal to a compression cuff and a magnetic tape recorder for recording of onset and disappearance of Korotkoff sounds. It weighed ~ 2.5 kg and was obviously inconvenient for an ambulatory patient to use. The first fully automatic device was developed, using compressed carbon dioxide to inflate the cuff. An electronic pump was introduced later and automatic data recording systems have been used since 1979.

Since then, the techniques for ambulatory blood pressure monitoring have been improved significantly. The modern device is light-weighted, compact in size, accurate, and automated in nature. It can be belt-worn and battery powered. The newest generation available in the current market is fully automatic, microprocessor-controlled, digitalized in memory, and extremely light weight (< 500 g).

Basic Techniques

The techniques for ambulatory blood pressure monitoring include auscultation, cuff oscillometry, and volume oscillometry.

Auscultation is a technique based on detection of onset and disappearance of Korotkoff sounds via a piezoelectric microphone taped over an artery distal to a deflating compression cuff. The Korotkoff sound is produced by turbulent flow while arterial blood flows through a segment of artery narrowed by a blood pressure cuff. The pressure at the onset of sound corresponds to systolic blood pressure, and at the disappearance of the sound to diastolic pressure. The advantage of this technique is simplicity, but the device is sensitive to background noise. This technique may also underestimate systolic pressure due to his flow dependency.

Cuff oscillometry is a technique based on detection of cuff pressure oscillations or vibrations to calculate systolic and diastolic values using an algorithmic approach. The systolic pressure corresponds to the cuff pressure at which oscillations first increase, and the diastolic pressure corresponds to the cuff pressure at which oscillations cease to decrease. The endpoints are estimated by analysis of oscillation amplitudes and cuff pressures. Different algorithms are used by different manufacturers, which may result in variability among different devices. This technique is insensitive to background noise, but arm movement may cause an errant reading. It may overestimate systolic pressure because of transmitted cuff pressure oscillations.

Volumetric oscillometry is a technique based on detection of finger volume pulsations under a cuff. The pressures are estimated as the cuff pressures at which finger volume oscillations begin (systolic pressure) and become maximal (mean pressure). Diastolic pressure is then derived from the known systolic and mean pressures. One problem with this technique is that this finger pressure may have a variable relationship to the brachial pressure. Another problem is that the technique cannot directly assess diastolic pressure.

Despite some problems associated with the mentioned techniques, their accuracy has been confirmed by validation testing using mercury sphygmomanometry and intraarterial measurement. The discrepancy is generally < 5 mmHg (0.399 kPa) between ambulatory devices and readings taken by trained professionals.

Patients are advised to wear the monitor for a period of 24 h, preferably during a normal working day. The monitor is preprogrammed to measure and record blood pressure at certain time intervals, preferably every 15–20 min during daytime hours and every 20–30 min during nighttime hours. Patients are also advised to document their activity during the testing period for assessment of any stress-related blood pressure.

The monitoring device consists of a small central unit and an attached cuff. The central unit contains a pump for cuff inflation and deflation, and the memory device, such as tape or digital chip, for recording. The time intervals between the measurements, maximal and minimal inflation pressures, and deflation rate are programmable according to the physician's order. The recording pressures can be retrieved from the tape or memory chip for analysis. Due to recent applications of digital technology and advanced software programs, a large amount of data can be stored in a small chip, and analysis can also be done automatically to generate a patient's report for the physician's use. A complete patient's report normally contains all blood pressure readings over a 24 h period, heart rates, mean arterial pressures, and statistic summaries for daytime, nighttime, and 24 h periods.

New Clinical Concepts Related to Ambulatory Blood Pressure Monitoring

A few new considerations related to ambulatory blood pressure monitoring have emerged. These include blood pressure load, pressure dipping, pressure variability, and white-coat hypertension. Health professionals need to understand these concepts in order to properly interpret or use data collected from monitoring.

Blood Pressure Load. This is defined as the proportion of the 24 h pressure recordings above the thresholds for waking and sleep blood pressure. The threshold commonly used for estimating the pressure load during waking hours is 140/90 and 120/80 mmHg (15.99/10.66 kPa) during sleep. Blood pressure load is helpful in the diagnosis of hypertension and in the prediction of end-organ damage. It has been considered closely correlated with left ventricle

hypertrophy. It has been reported that the incidence of left ventricular hypertrophy is $\sim 90\%$ in untreated patients with systolic blood pressure loads $> 50\%$, and $\sim 70\%$ with diastolic blood pressure loads $< 40\%$ (6,7).

Dipping and Circadian Blood Pressure Variability. Dipping is a term used to describe the circadian blood pressure variation during 24 h ambulatory blood pressure monitoring. In normotensive patients there is circadian blood pressure variability. Typically, the peak blood pressures occur around 6 a.m., and then taper to lower levels during the evening hours and further at night with the lowest levels between 2 and 4 a.m.. A patient whose blood pressure drops by at least 10% during sleep is considered normal (a dipper), and by $< 10\%$ abnormal (nondipper). In comparison to dippers, nondippers have been reported associated with higher prevalence of left ventricular hypertrophy, albuminuria, peripheral arterial changes, and cerebral lacunae. Nondippers have also been reported to have increased cardiovascular mortality rates (8).

White-Coat Hypertension. This is a condition in which blood pressure is persistently elevated in the presence of a doctor, but falls to normal levels when the patient leaves the medical facilities. Measurement by a nurse or trained nonmedical staff may reduce this effect. Because decisions regarding treating hypertension are usually made on the basis of isolated office blood pressure reading, a doctor may incorrectly diagnose this group of patients as sustained hypertension and prematurely start the therapy. This phenomenon has been reported in 15–35% of patients currently diagnosed and treated as hypertensive. However, white-coat hypertension can be easily detected by either ambulatory blood pressure monitoring or self-monitoring at home. It may or may not be benign, requiring definitive outcome studies to rule out any end-organ damages. It also requires continued surveillance by self-monitoring at home and repeat ambulatory blood pressure monitoring every 1–2 years (9,10).

Interpretation of Ambulatory Blood Pressure Profile

Normal ambulatory blood pressure values for adults are currently defined to be $< 135/85$ mmHg (17.99/11.33 kPa) during the day, $< 120/75$ mmHg (15.99/9.99 kPa) during the night, and $< 130/80$ mmHg (17.33/10.66 kPa) over 24 h. Daytime and night time blood pressure loads should be less 20% above normal values. Mean day-time and nighttime (sleep) blood pressure measurements should differ by at least 10%. The ambulatory blood pressure profile should also be inspected in relation to diary data and time of drug therapy.

Indications of Ambulatory Blood Pressure Monitoring

Although ambulatory blood pressure monitoring was originally developed as a research tool, it has widely been applied in clinical practice to help diagnose and manage hypertensive patients. It is indicated to rule out white-coat hypertension, to evaluate drug-resistant hypertension, to assess symptomatic hypertension or hypotension, to diagnose hypertension in pregnancy, and to assess adequacy of blood pressure control in patients at high risk of cardiovascular diseases.

White-Coat Hypertension. Office-based blood pressure measurement cannot differentiate sustained hypertension from white-coat hypertension. Historical appraisal and review of self-recorded blood pressures may aid in identification of patients with white-coat hypertension. However, ambulatory blood pressure monitoring is more effective in this clinical scenario to rule out white-coat hypertension. Recognition and proper management of patients with white-coat hypertension may result in a reduction in medication use and eliminate related cost and side effects. Although white coat hypertension may be a prehypertensive state and can eventually evolve to sustained hypertension, data collected from ambulatory blood pressure monitoring suggest, patients with white coat hypertension who maintain low ambulatory blood pressures (< 130–$135/80$ mmHg) have a low cardiovascular risk status and no demonstrable end-organ damage (11).

Drug-Resistant Hypertension. Drug resistant hypertension is defined as a condition when adequate blood pressure control ($< 140/90$ mmHg) (18.66/11.99 kPa) cannot be achieved despite the use of appropriately combined antihypertensive therapies in proper dosages for a sufficient duration. Ambulatory blood pressure monitoring helps evaluate whether additional therapy is needed. The causes include true drug-resistant hypertension as well as other conditions such as superimposition of white-coat hypertension on existing hypertension, patient's noncompliance, pseudohypertension secondary to brachial artery calcification, and sleep apnea and other sleep disorders. Ambulatory blood pressure monitoring can help differentiate the true drug resistant hypertension from the above-mentioned conditions (12).

Episodic Hypertention. A single office-based measurement of blood pressure may or may not detect episodic hypertension as in pheochromocytoma. In this clinical scenario the 24 h ambulatory blood pressure monitoring is a useful diagnostic tool. It is indicated if a patient's symptoms or signs are suggestive of episodic hypertension (13).

Borderline or Labile Hypertension. Patients with borderline hypertension often demonstrate only some (but not all) elevated blood pressure readings in office-based measurement, 24 h ambulatory blood pressure monitoring can benefit these patients and provide a useful diagnostic information for physician's use (14).

Hypertension with End-Organ Damage. Patients who exhibit worsening of end-organ damage may suggest inadequate 24 h blood pressure control. Occasionally, those patients may demonstrate adequate blood pressure control based on the office-based measurements. In this condition, a 24 h blood pressure monitoring is needed to rule out inadequate blood pressure control, which is associated with worsening of end-organ damage (15).

Hypentensive Patients with High Risk of Cardiovascular Events. Some hypertensive patients are at particularly high

risk of cardiovascular events, such as those with diabetes and/or past stroke. Those patients require rigorous blood pressure control over 24 h. Ambulatory blood pressure monitoring can be applied to assess the 24 h control (15).

Suspected Syncope or Orthostatic Hypotension. Transient hypotensive episodes and syncope are difficult to assess with the office-based blood pressure measurements, but are readily recorded with ambulatory blood pressure monitoring. Therefore, if symptoms and signs are suggestive of syncope or orthostatic hypertension, patients can benefit from 24 h blood pressure monitoring, especially in conjuction with Holter monitoring (15).

Hypertension in Pregnancy. About 10% of pregnancies may be complicated by hypertension. At the same time, white-coat hypertension may affect up to 30% of patients. It is important to differentiate true hypertension in pregnancy from white-coat hypertension, to avoid unwarranted hospitalizations or medication use. In this clinical scenario, ambulatory blood pressure monitoring would help to rule out white-coat hypertension and identify pregnancy-induced hypertension (16).

Clinical Research. Since ambulatory blood pressure monitoring can provide more samples of blood pressure measurements, data from this device is therefore much more statistically significant than a single isolated office-based reading. Therefore, statistical significance of clinical studies can possibly be achieved with smaller numbers of patients. This is very important for the efficient study of new therapeutic agents (17).

Limitations of Ambulatory Blood Pressure Monitoring

Although ambulatory blood pressure monitoring has been proved useful in the diagnosis and management of hypertension, the technology remains underused secondary to lack of experience in interpretation of results, unfamiliarity with devices, and some economic issues. Adequate staff training, regular calibration of devices, and good quality control are required. The patient's diary of daily activities and time of drug treatment are also needed for proper data analysis and interpretation.

Future Development

Like any other ambulatory device, an ideal noninvasive ambulatory blood pressure monitoring device should be user-friendly, light-weight, compact in size, digitalized for automated data management, and low in cost. Application of newer technologies will make such devices available, hopefully, in the near future.

AMBULATORY BLOOD GLUCOSE MONITORING

Introduction

Diabetes is one of most common diseases suffered by millions of people around the world. It is essential to monitor blood glucose to ensure overall adequate blood glucose control. Traditional standard blood glucose monitoring devices require invasive blood samplings and are therefore unsuitable for ambulatory blood glucose monitoring. Development of minimally invasive or noninvasive ambulatory glucose monitoring devices that provide accurate, near-continuous measurements of blood glucose level have the potential to improve diabetes care significantly. Such devices will provide information on blood glucose levels, as well as rate and direction of change, which can be displayed to patients in real-time and be stored for later analysis by physicians. Guardian RT system recently developed by Medtronic MiniMed is an example (Fig. 3). It provides continuous real-time glucose readings around the clock. Due to the huge market potential, many biomedical and medical instrument companies are developing similar devices for ambulatory glucose monitoring. Several innovative devices have recently been unveiled; many more are still in development. It is expected that some of them will be eventually U. S. Food and Drug Administration (FDA) approved as a replacement for standard blood glucose monitors, providing patients with a new option for long-term, daily monitors in the near future. The FDA is concerned about the accuracy of ambulatory continuous glucose monitoring devices when compared to the accuracy of standard monitoring devices. This issue will be eventually eliminated as related technologies become more and more mature. Technically, a typical ambulatory glucose monitoring device consists of a glucose sensor to measure glucose levels and a memory chip to record data information.

Figure 3. Ambulatory glucose monitoring with guardian real time system (Medtronic MiniMed).

Glucose Sensors

The glucose sensors for ambulatory glucose monitoring devices are either minimally invasive or completely noninvasive. A variety of technologies have emerged over the past decade aiming at development of ideal glucose sensors suitable for ambulatory monitoring.

A typical minimally invasive ambulatory continuous glucose sensor is a subcutaneous device developed by Minimed, Inc. (18). The sensor is designed to be inserted into a patient's abdominal subcutaneous tissue. It measures glucose levels every 10 s and records means > 5 min intervals. The technology involves measurement of glucose levels of

interstitial fluid via the subcutaneous sensor. The blood glucose levels are then derived from the measured interstitial fluid glucose levels. The detection mechanism involves use of a low fluorescence molecule. Electrons are transferred from one part of the molecule to another when excited by light. This prevents bright fluorescence from occurring (19). When bound to glucose, the molecule prevents the electrons from interfering with fluorescence, and the molecule becomes a bright fluorescent emitter. Therefore, the glucose levels can be determined based on the brightness of fluorescence. The glucose information will be transmitted from the sensor to a watch-like device worn on the wrist. Using this type of sensor, two devices have been developed by the company. One is a device that can be worn by the patient for a few days to record the glucose levels for the physician's analysis. The other is a device that can alert patients of impending hyperglycemia or hypoglycemia if the glucose levels go beyond the physician's predetermined upper and lower limits. The sensor can also work in conjunction with an implanted insulin pump, creating a "biomechanical" or artificial pancreas in response to the change at the glucose levels (20). It is predictable that such a biomechanical pancreas will eventually benefit millions of diabetic patients whose glucose control is dependant on insulin.

Complete noninvasive sensors for ambulatory glucose monitoring are even more attractive since they do not need any blood or interstitial samples to determine glucose levels. Several such sensors have recently been developed based on different technologies. For example, a glucose sensor that can be worn like a wristwatch has been developed by Pendragon Medical AG (Zurich, Switzerland). This sensor can continuously monitor blood glucose level without the need for a blood sample. It is based on impedance spectroscopy technology (21). The principle of this technology relates to the fact that blood glucose changes produce significant conductivity changes, causing electric polarization of cell membranes. At the same time, the sensor generates an electronic field that fluctuates according to the electrical conductivity of the body. A micro antenna in the sensor then detects these changes and correlates them with changes in serum glucose. With this technology, blood glucose levels can be monitored noninvasively in real time. Another promising noninvasive sensor is based on the possibility of measuring glucose by detecting small changes in the retinal capillaries. By scanning the retinal microvasculature, the sensor can directly measure glucose levels in aqueous humor using a reflectometer. Recently, a plastic thin sensor, which can be worn like a contact lens, has been innovated (22,23). The sensor changes its color based on the concentration of glucose, from red, which indicates dangerously low glucose levels, to violet, which indicates dangerously high glucose concentrations. When glucose concentration is normal, the sensor is green. Integration of the sensor material into commercial contact lenses may also be possible with this technology.

Memory Chips

Memory chips are used to record glucose data information for later use by the physician. The digital chips have many advantages, such as compact size, large memory, easy data transmission via wire or wireless, and possible autoanalysis with computer software. Patients can also upload their glucose data from digital memory chips to web-based data management systems, allowing diabetic patients and their health care providers to analyze and communicate glucose information using the internet.

Significances of Ambulatory Glucose Monitoring

Ambulatory glucose monitoring can provide continuous data on blood glucose levels. Such data can improve diabetic care by enabling patients to adjust insulin delivery according to the rate and direction of blood glucose change, and by warning of impending hypoglycemia and hyperglycemia. Doctors can use ambulatory glucose monitoring to help diagnose problematic cases, fine-tune medications, and get tighter control of blood glucose levels for high risk patients. Obviously, the monitoring will improve overall blood glucose control, reducing short-term adverse complications and delaying onset of long-term serious complications, such as end-stage renal disease, heart attack, blindness, stroke, neuropathy, and lower extremity amputation.

In addition, continuous ambulatory glucose monitoring is a key step toward the development of artificial pancreas, which could deliver insulin automatically in response to blood glucose levels. It is expected that such an artificial pancreas would greatly benefit many diabetic patients and provide them new hope for better quality of life.

Future Development

Although many continuous ambulatory glucose monitoring devices are still in the stage of clinical trials, there is little doubt as to the value of the devices in management of diabetic patients. It is expected that millions of diabetic patients will be benefited once such devices are widely available. At the same time, introduction of more and more new devices highlights the need for careful evaluation to ensure accuracy and reliability. Cooperation between the manufacturers and physicians to fine-tune the technology will eventually lead to approval of the devices by the FDA to replace traditional invasive standard glucose monitoring. Technology for continuous ambulatory glucose monitoring is also required to make an artificial pancreas, which would offer great hope for millions of patients with diabetes.

CONCLUSION

Ambulatory monitoring has increasingly provided a powerful alternative tool to diagnose and manage some diseases. Continuous advancement in a variety of technologies provides more and more innovative ambulatory devices to serve the patients' need. Applications of information technology and specialized software tools make autotransmission and autoanalysis of ambulatory monitoring data possible. Clinicians will be able to monitor their ambulatory patients distantly without a hospital or office visits. In addition, integration of the technology of continuous ambulatory monitoring with an implantable automatic therapeutic pump may create a biomechanical system in response to specific abnormal changes. The artificial pancreas currently in development is a typical example for

such hybrid devices. Such devices will be available in the market in the near future.

BIBLIOGRAPHY

Cited References

1. Holter NJ. New method for heart studies: Continuous electrocardiography of active subjects over long periods is now practical. Science 1961;134:1214–1220.
2. Heilbron EL. Advances in modern electrocardiographic equipment for long-term ambulatory monitoring. Card Electrophysiol Rev 2002;6(3):185–189.
3. Kadish AH, et al. ACC/AHA clinical competence statement on electrocardiography and ambulatory electrocardiography: a report of the ACC/AHA/ACPASIM task force on clinical competence. Circulation 2001;104:3169–3178.
4. Kinlay S, et al. Event recorders yield more diagnoses and are more cost-effective than 48 hour Holter monitoring in patients with palpitations. Ann Intern Med 1996;124:16–20.
5. Hinman AT, Engel BT, Bickford AF. Portable blood pressure recorder accuracy and preliminary use in evaluation intradaily variations in pressure. Am Heart J 1962;63:663–668.
6. Zachariah PK, et al. Blood pressure load: A better determinant of hypertension. Mayo Clin Proc 1998;63:1085–1091.
7. White WB, Dey HM, Schulman P. Assessment of the daily blood pressure load as a determinant of cardiac function in patients with mild-to-moderate hypertension. Am Heart J 1989;118:782–795.
8. Pickering TG. The clinical significance of diurnal blood pressure variations: dippers and nondippers. Circulation 1990;81:700–702.
9. Verdecchia P, et al. White-coat hypertension: not guilty when correctly defined. Blood Press Monit 1998;3:147–152.
10. Pickering TG, et al. How common is white coat hypertension. Hypertension 1988;259:225–228.
11. Palatini P, et al. Target-organ damage in stage-1 hypertensive subjects with white coat and sustained hypertension: results from the HARVEST study. Hypertension 1998;31:57–63.
12. Brown MA, Buddle ML, Martin A. Is resistant hypertension really resistant ? Am J Hypertens 2001;14:1263–1269.
13. Myers MG, Haynes RB, Rabkin SW. Canadian hypertension society quidelines for ambulatory blood pressure monitoring. Am J Hypertens 1999;12:319–331.
14. Pickering T. for the American Society of Hypertension ad-hoc Panel. Recommendations for the use of home (self) and ambulatory blood pressure monitoring. Am J Hypertens 1996;9:1–11.
15. O'Brien E, et al. Use and interpretation of ambulatory blood pressure monitoring: recommendations of the British Hypertension Society. BMJ 2000;320:1128–1134.
16. Halligan A, et al. Twenty-four-hour ambulatory blood pressure measurement in a primigravid population. J Hypertens 1993;11:869–873.
17. Conway J, et al. The use of ambulatory blood pressure monitoring to improve the accuracy and reduce the numbers of subjects in the clinical trials of antihypertensive agents. J Clin Exper Hypertension 1986;8:1247–1249.
18. Cross TM, et al. Performance evaluation of the MinMed continuous glucose monitoring system during patient home use. Diab Technol Ther 2000;2:49–56.
19. Pickup JC, Shaw GS, Claremont DJ. *In vivo* molecular sensing in diabetes mellitus: an implantable glucose sensor with direct electron transfer. Diabetes 1989;32:213–217.
20. Jaremko J, Rorstad O. Advances toward the implantable artificial pancreas for treatment of diabetes. Diabs Care 1998;21:444–450.
21. Caduff A, et al. First human experiments with a novel noninvasive, non-optical continuous glucose monitoring system. Biosens Bioelecs 2003;19:209–217.
22. Badugu R, Lakowicz JR, Geddes CD. Ophthalmic glucose sensing: a novel monosaccharide sensing disposable and colorless contact lens. Analyst (England) 2004;129:516–521.
23. Badugu R, Lakowicz JR, Geddes CD. Ophthalmic glucose monitoring using disposable contact lenses—a review. J Fluoresc 2004;14:617–633.

See also ARRHYTHMIA ANALYSIS, AUTOMATED; BIOTELEMETRY; HOME HEALTH CARE DEVICES; PACEMAKERS.

ANALYTICAL METHODS, AUTOMATED

LAKSHMI RAMANATHAN
Mount Sinai Medical Center
LASZLO SARKOZI
Mount Sinai School of Medicine

INTRODUCTION

The chemical composition of blood, urine, spinal fluid, sweat, provides a wealth of information on the well being or illness of the individual. The presence, concentration, and activity of chemical constituents are indicators of various organ functions. Concentrations higher or lower than expected sometimes require immediate attention. Some of the reasons to analyze body fluids:

1. Screening of an apparently healthy population for unsuspected abnormalities.
2. Confirming or ruling out a diagnosis.
3. Monitoring changes during treatment, improvement of condition or lack of improvement.
4. Detecting or monitoring drug levels for diagnosis or maintenance of optimal therapeutic levels.

By the 1950s, demands of clinicians for laboratory tests increased rapidly. Classical methods of manual laboratory techniques could not keep up with these demands. The cost of performing large numbers of laboratory tests by manual methods became staggering and the response time was unacceptable.

The article in the first edition of this Encyclopedia published in 1988 describes the history of laboratory instrumentation during the previous three decades (1). Reviewing that long list of automated instruments, with the exception of a few, all became museum pieces. During the last 15 years the laboratory landscape changed drastically. In addition, new group of automated instruments were introduced during this period. They were developed to perform bedside or near patient testing, collectively called Point of Care Testing instruments. In this period in addition to new testing instruments, perianalytical instrumentation for specimen handling became available. Their combined result is increased productivity and reduction of manpower requirements, which became imperative due to increased cost of healthcare and dwindling resources.

This article will present some financial justification of these investments.

PATIENT PREPARATION, SPECIMEN COLLECTION, AND HANDLING

The prerequisites for accurate testing include proper patient preparation, specimen collection, and specimen handling. Blood specimens yield the most information about the clinical status of the patient though in many cases urine is the preferred sample. For specialized tests, other body fluids that include sweat and spinal fluid are used. When some tests, such as glucose and lipids, require fasting specimens, patients are prepared accordingly.

Common errors affecting all specimens include the following:

- Inaccurate and incomplete patient instructions prior to collection.
- Wrong container/tube used for the collection.
- Failure to label a specimen correctly.
- Insufficient amount of specimen to perform the test.
- Specimen leakage in transit due to failure to tighten specimen container lids.
- Interference by cellular elements of blood.

Phlebotomy techniques for blood collection have considerably improved with better gauge needles and vacuum tubes for collection. The collection tubes are color coded with different preservatives so that the proper container can be used for a particular analyte. The cells should be separated from the serum by centrifugation within 2 h of collection. Grossly or moderately hemolyzed specimens may be unsuitable for certain tests. If not separated from serum or plasma, blood cells metabolize glucose and produce a false decrease of ~5%/h in adults. The effect is much greater in neonates (2). If there is a delay in separating the cells from the serum, the blood should be collected in a gray top tube containing sodium fluoride as a preservative that inhibits glycolysis.

Urine collection is prone to errors as well, some of which include (3):

- Failure to obtain a clean catch specimen.
- Failure to obtain a complete 24 h collection/aliquot or other timed specimen.
- No preservative added if needed prior to the collection.

Once specimens are properly collected and received in the clinical laboratory, processing may include bar coding, centrifugation, aliquoting, testing and reporting of results.

AUTOMATED ANALYZERS

A large variety of instruments are available for the clinical chemistry laboratory. These may be classified in different ways based on the type of technology applied, the test menu, the manufacturer, and the intended application. Depending on the size of the laboratory, the level of automation varies. Clinical chemistry analyzers can be grouped according to throughput of tests and diversity of tests performed and by function, such as immunoassay analyzers, critical care blood gas analyzers, and urinalysis testing systems. Point of Care analyzers vary in terms of accuracy, diversity and menu selection.

Some of the features to consider while evaluating low or high volume analyzers are listed below:

- Test menu available on instrument:
  - Number of different measured assays onboard simultaneously.
  - Number of different assays programmed/calibrated at one time.
  - Number of user-defined (open) channels.
- Reagents:
  - Preparation of reagents if any.
  - Storage of reagents.
  - On board stability.
  - Bar-coding for inventory control.
- Specimen volume:
  - Minimum sample volume.
  - Dead volume.
- Instrument supplies:
  - Use of disposable cuvettes.
  - Washable/reusable cuvettes.
- Clot detection features along with quantitation of hemolysis and turbidity detection.
- Auto dilution capabilities of analyzer.
- Frequency of calibration.
- Quality control requirements.
- Stat capability.
- LIS interface.
- Maintenance procedures on instrument; anticipated downtime.
- Analyzer costs expressed in cost per reportable test.

Our goal is not to review every analyzer available on the market. We have chosen a few of the instruments–vendors. This is by no means endorsing any particular vendor, but merely discussing some of the most frequently utilized features or describing our personal experiences. The College of American Pathologists has provided excellent surveys of instruments and the reader is referred to those articles for more complete details (4).

CHEMISTRY ANALYZERS

Routine chemistry analyzers have broad menus capable of performing an average of 45 (20 to >70) different on board tests simultaneously, selected from an available menu of 26

Table 1. Automated Analyzers from Different Manufacturers

Instrument Type	Generic Menu	Vendor
Routine chemistry	Electrolytes, BUN, Glucose, Creatinine, Protein, Albumin, Lipids, Iron, Drugs of abuse, Therapeutic drug monitoring, etc.	Abbott, Bayer, Beckman Dade, J&J, Olympus, Roche
Immunoassays	Tumor markers, Cardiac markers, Anemia, B12, Folate and misc. Endocrine tests	Abbott, Bayer, Beckman, DPL, J&J, Olympus, Roche
Critical Care	Blood gases, Cooximetry, Electrolytes Ionized calcium, Lactate, Hematocrit	Abbott, Bayer, Instrumentation Lab, Nova, Radiometer, Roche

to >100 different analytes (5,6). Selection is based on test menu, analytic performance, cost (reagents, consumables and labor), instrument reliability (downtime etc.), throughput, and ease of use, customer support and robotic connectivity, if needed. Some automated analyzers from different manufacturers are listed in Table 1.

General Chemistry

Virtually all automated chemistry analyzers offer random access testing, multiple tests can be performed simultaneously and continuously. This is different from batch-mode instruments that perform a single test on a batch of samples loaded on the instrument (Abbott TDX and COBAS Bio). Many analyzers are so-called "open systems" that use reagents from either the instrument manufacturer or different vendors. The advantage of these systems being that the customer has a choice of reagent vendors and the reagent can be selected based on performance and cost.

An example of a closed system is a line of analyzers manufactured by Ortho Clinical Diagnostics. The Vitros 950 and the analyzers in this category use a unique, dry chemistry film-based technology developed by Kodak. The slide is a dry, multilayer, analytical element coated on a polyester support. A 10 μL drop of patient sample is deposited on the slide and is evenly distributed by the spreading layer to the underlying layers that contain the ingredients for a particular chemical reaction.

The reaction slide (Fig. 1.) for albumin shows the reactive ingredient is the dye (bromcresol green), which is in the reagent layer. The inactive ingredients that include polymeric beads, binders, buffer, and surfactants are in the spreading layer. When the specimen penetrates the reagent layer, the bromcresol green (BCG) diffuses to the spreading layer and binds to albumin from the sample. This binding results in a shift in wavelength of the reflectance maxima of the free dye. The color complex that forms is measured by reflectance spectrophotometry. The amount of albumin-bound dye is proportional to the concentration of albumin in the sample. Once the test is completed the slide is disposed into the waste container.

Some manufacturers close their system by labeling their individual reagent packs with unique barcodes, rejecting packs not distributed by them. Examples of "open systems" include analyzers manufactured by Olympus, Roche (Fig. 2.), Beckman, Dade and Abbott. Many instruments have both open and closed channels allowing greater flexibility in the use of reagents. In addition to diverse menus, open and closed channels, compatibility of analyzers

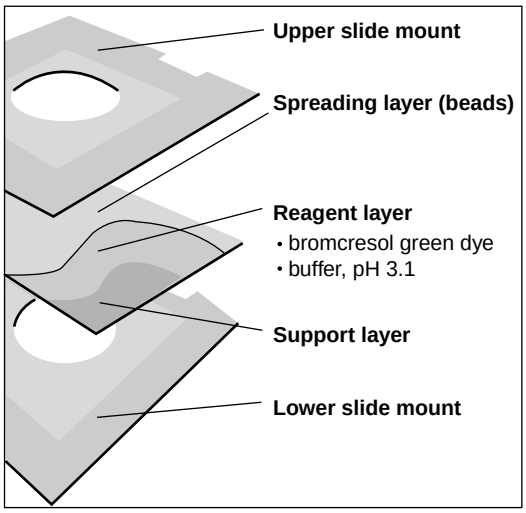


Figure 1. The Vitros 950 (J&J Diagnostics) slide is a dry, multilayer, analytical element coated on a polyester support. A drop of patient sample is deposited on the slide and is evenly distributed by the spreading layer to the underlying layers that contain the ingredients for a particular chemical reaction.

Figure 2. The Roche/Hitachi Modular™ analytic system has a theoretical throughput of 3500–5000 tests or 150–250 samples/h. They test 24 different analytes simultaneously with a total menu of > 140 available tests.

with perianalytical technology is becoming an important feature.

Perianalytical systems include front-end automation with specimen processing and aliquoting, track systems or other technologies to move specimens between instruments in the laboratory, and robots to place specimens on and remove them from the analyzers.

Immunoassay Analyzers

Immunoassay systems are presently the fastest growing areas of the clinical laboratory where advances in immunochemical methodology, signal detection systems, microcomputers and robotic processing are taking place at an accelerated pace (7). At present, manufacturers have high volume immunoassay analyzers that can be modularly integrated along with chemistry and hematology analyzers into fully automated laboratory systems. In addition, expanding menus of homogeneous immunoassays allow integration into many laboratories using "open reagent kits" designed for use on automated clinical chemistry analyzers.

One of the several analyzers in this category is the Bayer Advia Centaur (Fig. 3.)

Of the different enzyme immunoassays (EIA) available, only the two homogeneous methods, EMIT and CEDIA have been easily adapted to fully automated chemistry analyzers (8–11). The other EIAs require a separation step to remove excess reagent that will interfere with the quantitation of the analyte. Abbott uses a competitive assay involving a fluorescent-labeled antigen that competes for a limited number of sites on antigen specific antibody. The amount of analyte is inversely proportional to the amount of fluorescence polarization. Chemiluminescence technology is used in the Bayer ACS and Roche Elecsys systems combines very high sensitivity with low levels of background interference. Essentially, it involves a sandwich immunoassay direct chemiluminometric technology, which uses constant amounts of two antibodies. The first antibody in the Lite Reagent is a polyclonal goat anticompound antibody labelled with acridinium ester. The second antibody in the Solid Phase is a monoclonal mouse anticompound antibody, which is covalently coupled to paramagnetic particles. A direct relationship exists between the amount of compound present and the amount of relative light units (RlU) detected by the system (Table 2).

Critical Care Analyzers

Blood gas measurements performed on arterial, venous, and capillary whole blood includes electrolytes and other tests in addition to the gases. These tests are listed in Table 3.

The Nova CCX series combines blood gas measurements with co-oximetry, electrolytes, a metabolic panel and hematology on 50 μL of whole blood. Several blood gas analyzers are utilizing the concept of "Intelligent Quality Management" whereby the analyzers run controls automatically at specified time intervals set by the operator. If a particular analyte is not within the specified range, the analyzer will not report out any patient results on the questionable test. Selected blood gas and critical care analyzers are listed in Table 4.

The unique specimen and turnaround time requirements for blood gases have prevented the tests from

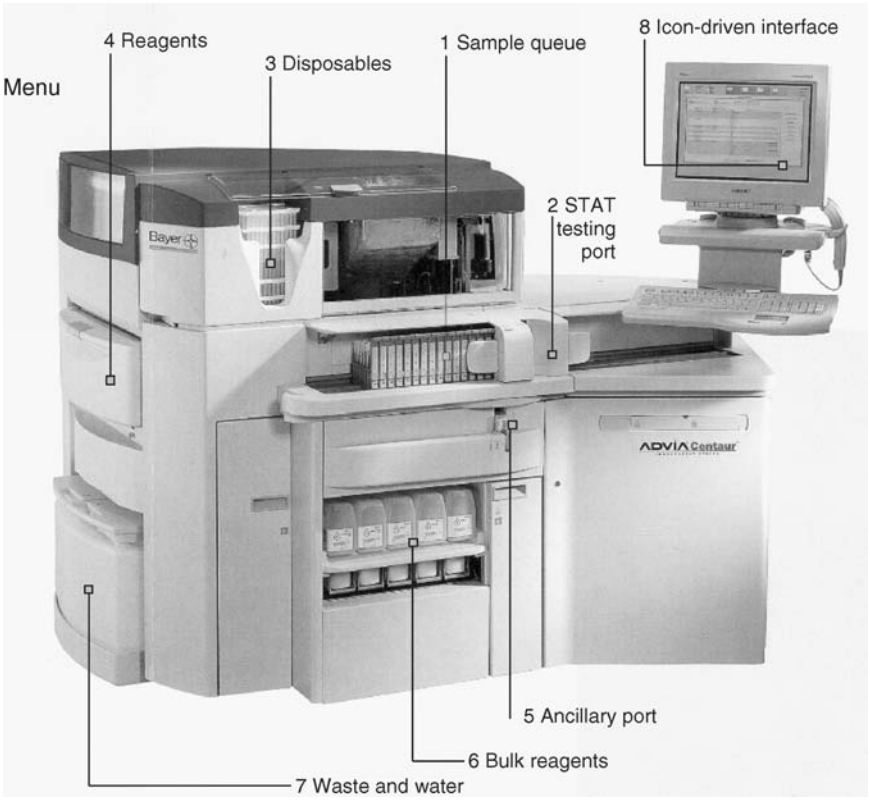


Figure 3. The Bayer Advia Centour system has large on-board capacity for reagents and supplies combined with automated maintenance and monitoring features streamline operations. Categories such as fertility, therapeutic drug monitoring, infectious disease, allergy, cardiovascular, anemia, and oncology, therapeutic drug monitoring and thyroid tests are available. Up to 30 different reagent packs can be placed on the instrument. It has a throughput of 240 tests/h.

Table 2. Immunoassay Analyzers

Manufacturer	Model	Methodology
Abbott diagnostics	Axsym	FPIA, MEIA
	TDX, IMX	FPIA
	ADX	FPIA
	Architect	Chemiluminescence
Bayer Diagnostics	ACS 180	Chemiluminescence
	Centaur	Chemiluminescence
	Immuno 1	EIA
Beckman Coulter	Access	EMIT
	LX-20	EMIT
	DCI	Chemiluminescence
Boehringer Manheim ES-300	EIA	
	Elecsys	Chemiluminescence
Dade Behring	Opus Magnum	EIA
	Stratus	FIA
	ACA	EIA, Petinia
	ACA	EIA, turbidimetric
Diagnostic Product Corp.	Immulite	Chemiluminescence, EIA
Nichols Diagnostics CLS ID	Chemiluminescence	
Ortho Clinical	Eci	Chemiluminescence

Table 3. Test Menus for Critical Care Analyzers

Category	Tests Included
Blood gases	pH, pCO_2, pO_2 and other calculated parameters
Electrolytes	Sodium, potassium, chloride, bicarbonate, ionized calcium
Co-oximetry	Carboxyhemoglobin, methemoglobin, total hemoglobin, O_2 saturation
Metabolic panel	Glucose, blood urea nitrogen, creatinine, lactate
Hematology	Hematocrit, hemoglobin, activated clotting time

being performed in combination with general chemistry tests.

Point of Care Testing

Point of care testing (POCT) is defined as laboratory diagnostic testing performed close to the location of the patient. Recent advances over the last decade have resulted in smaller, more accurate devices with a wide menu of tests (12,13). Today POCT can be found from competitive sports to the prison system, from psychiatric counseling to preemployment and shopping mall health screening. Use of POCT devices can be found in mobile transport vehicles such as ambulances, helicopters cruise ships and even the space shuttle.

The advantage of POCT is the ability to obtain extremely rapid laboratory results. However, it is necessary to be aware of the limitations of POCT devices in clinical practice. Venous blood samples often have to be drawn and sent to the main laboratory for confirmation if the results are not within a certain specified range. Another disadvantage of POCT is costs.

In compliance with the guidelines set by federal, state regulatory agencies and the College of American Pathologists (CAP), point of care testing programs are usually overseen by dedicated staff under the direction of the central laboratory. The responsibilities of the POCT staff include education and training of hospital staff, troubleshooting of equipment, maintaining quality control and

Table 4. Partial List of Critical Care Instruments

Vendor	Instrument
Abbott (iSTAT)	iSTAT
Bayer	200,300 800 series, Rapidpoint
Diametrics	IRMA
Instrumentation Lab	1600, 1700 series, Gem series
NOVA	Stat profile series, CCX
Radiometer	ABL series
Roche (AVL)	900 series, Omni and Opti series

Figure 4. The Roche Accu-Check is a small, easy to use blood glucose meter; it is widely used by our Point of Care Testing program. Test results are downloaded to the Laboratory Information System.

quality assurance standards. For a successful POCT program, the laboratory and clinical staff need to effectively work together.

The handheld Accu-Chek POCT device is shown on Fig. 4.

The most widely used point of care tests are bedside glucose testing, critical care analysis, urinalysis, coagulation, occult blood and urine pregnancy testing. Selected point of care devices are listed in Table 5. Other available POCT tests: cardiac markers, pregnancy, influenza A/B, Rapid Strep A, *Helicobacter pylori*, urine microalbumin and creatinine.

CLINICAL LABORATORY AUTOMATION

Historical Perspective

Along with innovations in instrumentation, automating perianalytical activities such as centrifuging, aliquoting, delivering specimens to the automated testing instruments, recapping and storing plays significant role in the modern clinical laboratory (14). Robotic systems that automate some or virtually all of the above functions are available. Automated laboratory and information systems offer benefits in terms of speed, operating efficiency, integrated information sharing and reduction of error.

Table 5. Selected Point of Care Devices

Test	Vendor
Bedside glucose test	Abbott (Medisense PCx)
	Bayer
	Ortho (Lifescan: One Touch)
	Roche
Critical care	Abbott (iSTAT)
	Bayer (Rapidpoint)
	IL (Gem series)
Coagulation	Abbott (iSTAT)
	Bayer (Rapid point)
	Hemosense
	ITC (Hemochron series)
	Medtronics (Hepcon)
	Roche (Coaguchek)
Fecal occult blood	Helena
	Smithkline Diagnostics
Urinalysis	Bayer (Multistix and Clintek)
	Roche (Chemstrip and CUA)

However, the individual needs of each laboratory have to be considered in order to select the optimum combination of instrumentation and perianalytical automation. For small laboratories, front-end work cell automation may be applied economically. For large commercial reference labs and hospital labs, total laboratory automation (TLA) is appropriate where samples move around the whole lab, or from place to place (15).

Clinical laboratory automation evolved with the development of the hematology "Coulter Counter" and the chemistry "AutoAnalyzer" in the 1950s. Automated cell counting by the Coulter involved placing a sample of whole blood in a hemocytometer and using a microscope to count the serial passage of individual cells through an aperture. Likewise, the automated analysis of patient samples for several chemistries dramatically changed the testing process in the chemistry laboratory.

In the 1980s in Japan, Dr. Sasaki's group developed a point-to-point laboratory system that was based on overhead conveyor transportation, delivering specimens placed in 10 position racks (16). These initial designs are the basis of several automation systems available today.

Automation Options and System Design

Available options for automation include the following:

- Interfaced instruments (some can be operated as stand alone analyzers and later linked to a modular system).
- Modular instruments (including, processing, and instrument work cells).
- Multidiscipline platforms (including multifunction instruments and multiwork cells).
- Total laboratory automation robotics system that automates virtually all routine functions in the laboratory.

Automation system design usually rests on the needs of the user. However, the following concepts should be considered:

- Modern information technology with hardware and operating systems that are vertically upgraded.
- Transportation system management at both the local level (device) and overall system level.
- Specimen tracking so that any specimen can be located in the automation system.
- Reflex testing where an additional test can be performed at the same instrument or the specimen can be retrieved to another instrument.
- Information systems agreement with the Laboratory Information System (LIS).

The ability to interface between the hospital LIS and the laboratory automation system (LAS) has been significantly

enhanced by the implementation of the HL7 system-to-system interface. The National Committee on Clinical Laboratory Standards (NCCLS) has issued a proposal level standard (Auto 3 –P) that specifies the HLA interface as the system-to-system communications methodology for connecting an LIS and an LAS (17–21).

NCCLS Guidelines

Components of an optimized laboratory automation system per NCCLS may include:

Preprocessing specimen sorting.
Automated specimen centrifugation.
Automated specimen aliquoting.
Specimen–aliquot recapping/capping.
Specimen integrity monitoring.
Specimen transportation.
Automated specimen sampling.
Automated specimen storage and retrieval.

It is also recommended that process control software should support:

Specimen routing.
Reflex testing.
Repeat testing.
Rules based processing.
Patient data integration.

Available Automation Systems

In the mid-1990s, several laboratory automation technologies implemented hardware-based automation solutions that were centered on defining a limited number of specimen containers compatible with the transportation system. By limiting the number of specimen containers, the hardware can be better defined and more efficient. The original Coulter IDS automation system and the original Hitachi CLAS were based on fixed, rigid or hard-coded hardware technologies.

In the Hitachi CLAS and modular systems, the automation transportation devices use the Hitachi 747 five-place specimen container rack. In order to move the specimen container rack from one analyzer to the next, the automation system must carry along four other patient specimens. The requirement to carry along additional specimens along with the target specimen creates significant mathematical complexity in routing and scheduling of tests. The use of a simple specimen container per specimen carrier model allows the routing of an individual specimen to a workstation without interrupting the flow of other individual specimens in the system.

Total laboratory automation is used to describe the Beckman Coulter IDS system (22). We have two parallel systems in our laboratory (Fig. 5). The basic components include the inlet module, where samples are placed, a centrifuge, serum level sensor, decapping unit, aliquoter-labeler units, outlet units, refrigerated storage unit and a disposal unit. A line PC that interacts with the LIS and all the individual components of the automation system controls the entire system. Each of the automated instruments has their own individual attachment for the handling of specimens being received from the robotic system. View of our automated (perianalytical and analytical) clinical laboratory is shown on Fig. 6.

Work Cell Technologies

The work cell model can be divided into two basic approaches. The first includes all instruments from the same discipline (Chemistry). The second approach is the development of a platform that includes multiple disciplines. An example of this is the Bayer Advia work cell in which chemistry, hematology, immunoassay, and urinalysis processing can take place on one platform. However, this work cell does not have front-end specimen processing and handling capability. Several automated work cells are available in the market at the present time. They include Abbott (Abbott hematology work cell), Beckman-Coulter (Acel-Net work cell), Bayer (Advia work cell), Johnson and Johnson (lab interlink labframe select), and Roche (modular system). The work cell technology varies from simple specimen transportation to complex specimen management.

LABORATORY AUTOMATION-A FINANCIAL PERSPECTIVE

Several studies are being reported on the financial aspects of automation. The most significant impact has been the reduction in FTEs and improvement in turnaround time. A retrospective analysis of 36 years of the effects of initially automation followed by total laboratory automation in the clinical chemistry laboratory at Mount Sinai Medical Center indicated that workload was significantly increased with a reduction of personnel (23). We present these productivity changes in Table 6.

Increased productivity resulted in significant reduction of performing laboratory tests (Table 7).

The effect of increased productivity is illustrated by the drastic reduction of cost/test (Fig. 7)

CALCULATIONS FOR NET PRESENT VALUE OF THE MOUNT SINAI CHEMISTRY AUTOMATION PROJECT (FIG. 8)

Net Present Value

The Net Present Value (NPV) is the value of the net cash flows generated by the project in 1998 $ (the year in which the project was initiated). The NVP is calculated by discounting the value of the annual cash flows [using values taken from the Present Value Interest Factor (PVIF) table for a given project length and cost of capital] to the purchasing value of the dollar at the date of inception of the project (1998). The length of the investment project is a conservative estimate of the useful economic lifetime of the investment project. In this case, we believe that after 8 years additional investments in upgrades

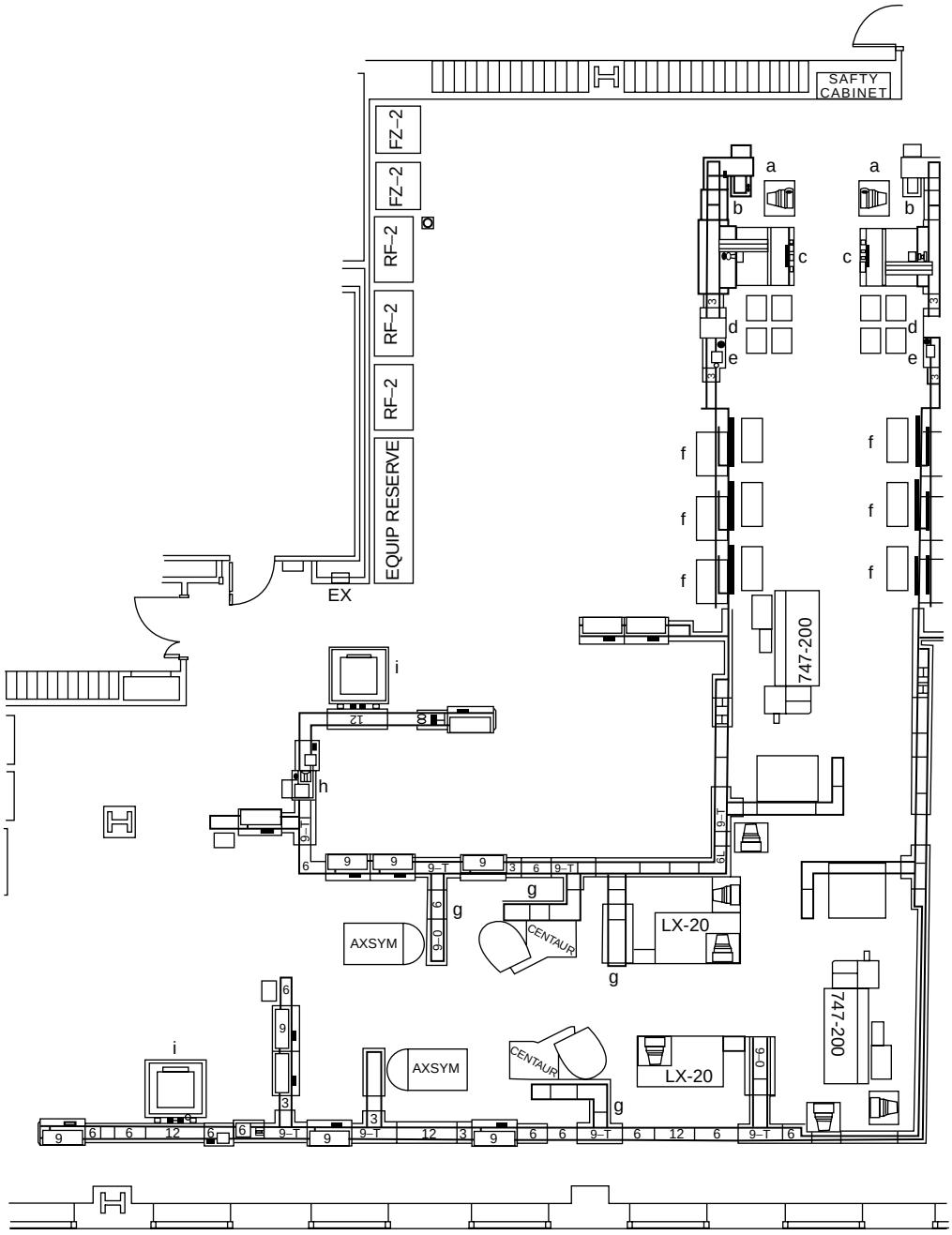


Figure 5. Floor plan of our Total Laboratory Automation. (a) Sample reception. Specimens are picked up, 5 at a time, from 50-position racks and loaded into individual tube holders. A bar-code verification unit determines the legibility of the labels, determines if the specimen is on the right processing location, rejects the suspect samples into a holding area and accepts the correct ones by a message to the Laboratory Information System:"Specimen Received". (b) Sample transport. The transport lanes are conveyor belts that move the samples about the system. (c) Centrifugation. Samples are loaded and unloaded automatically. The rotor has 50 positions. In our laboratory 350 specimens/h can be processed in these centrifuges. (d) Serum level detection. After centrifugation the samples are lowered into an optical well and based on the transmitted information the amount of the available serum is calculated. (e) Cap removal. A gentle rocking motion removes the cups without creating potentially hazardous aerosols. (f) Tube labeling and sample aliquoting. For each primary serum sample secondary aliquot tuber are prepared. The tube labeler prints a bar-code label and applies it to each aliquot tube. The number of aliquot tubes is defined by the system. Disposable pipette tips transfer the serum from the primary to the secondary (aliquot) tubes. The primary tubes are directed to a storage unit. (g) Instrument connections. Several instruments are connected to the transport system. Connection units load and unload samples. Samples not going to the analyzer can continue down the main line. (h) Cap replacement. When the testing of a secondary aliquot tube has been completed, the tube is directed toward an outlet unit, stockyard or storage locker. Before storage, the tube can receive a clean cap. (i) Refrigerated Sample storage. It holds up to 3000 tubes. Samples can be retrieved automatically through a request in the computer and sent to the location requested by the operator.

Figure 6. A portion of the Chemistry automated Core Laboratory at The Mount Sinai Hospital, New York.

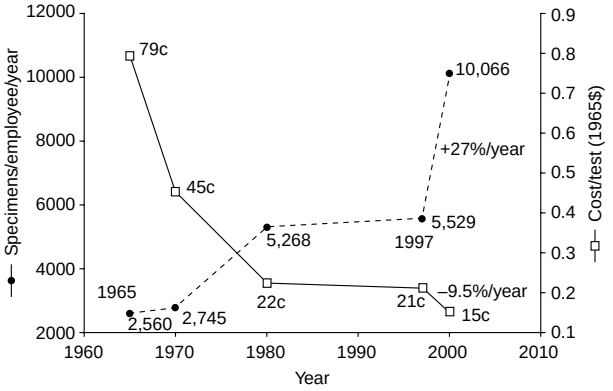


Figure 7. Automation increased Productivity and reduced cost. While the number of specimens processed increased from 2.500 to 10,000 specimens/year the cost/test was reduced from $0.79 to 0.15 (in 1965$).

beyond normal maintenance may be required. The cost of capital used was the interest rate of the lease taken out to finance the project. The relevant calculations are shown below:

1. Total cost of the lease (capital and interest):	$121,963/month×60 months =	$7,318,480
2. Total interest paid over the life of the lease:	$7,318,380 − $6,194,650 =	$1,123,730
3. Annual interest payments:	$1,123,730/5 years =	$ 224,730
4. Interest rate paid on lease:	($224,746/$6,194,650) × 100 =	**3,628%**

Negative Cash Flows

Negative cash flows represent money spent on the project. This includes capital outlays, lease payments ($3,140,000 or $741,921/year for 5 years, represented the portion on chemistry automation), project-related expenses (annual maintenance contract, years 1999 and 2000 = $74,240 annually, 2001–2005 = $39,000 annually.

Positive Cash Flows

Positive cash flows are those that represent money saved and/or costs avoided as the result of the chemistry automation project. There are recurring positive cash flows, resulting from savings that are essentially perpetual, such as salaries and benefits of workers replaced permanently by the chemistry automation project. Savings realized in a given year that are not expected to be repeated in subsequent years are nonrecurring positive cash flows. Staff pay raises during the years 1998, 1999, 2000, and 2002 were

Table 6. Increased Productivity

Year	Tech Staff	Other Staff	Total Staff	No. of Tests/Tech	No. of Tests/tot. Staff	No. of Tests/Specimen	Total No. of Specimens
1965	19	6.00	24.00	14,000	10,600	4.2	2,560
1970	34	17.00	51.00	36,205	24,150	8.8	2,745
1980	39	22.00	61.00	82,359	53,732	10.0	5,268
1997	38	17.00	55.00	94,058	66,099	11.8	5,529
2000	29	13.00	42.00	151,190	104,558	10.4	10,066
2002	29	39.00	35.00	169,582	128,530	10.5	12,190

Table 7. Cost/Test Reduction

Year	Tech Salary, $	Salary $/Test	Supplies $/Test	Total $/Test	Salary 1965 $/Test	Supplies 1965 $/Test	Total 1965 $/Test
1965	5,170	0.70	0.19	0.79	0.70	0.09	0.79
1970	9,114	0.38	0.17	0.55	0.31	0.14	0.45
1980	16,500	0.37	0.20	0.57	0.14	0.08	0.22
1997	38,000	0.66	0.41	1.07	0.13	0.08	0.21
2000	41,000	0.45	0.36	0.81	0.08	0.07	0.15
2002	44,000	0.38	0.34	0.72	0.07	0.06	0.13

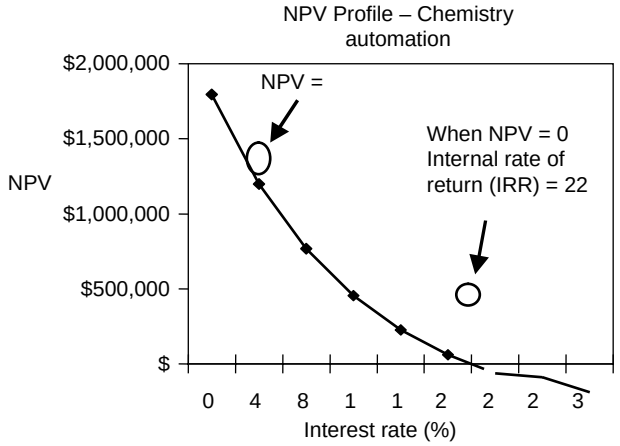


Figure 8. The Internal Rate of Return for the Chemistry Automation project was a remarkable 22%.

financed directly from chemistry automation project savings. These amounts are not reflected in the net salary and benefits savings. As such they are positive cash flow, since they represent costs covered by the automation savings that otherwise would have had to be financed through other sources.

The NPV Profile and Internal Rate of Return

The interest rate employed to discount the value of cash flows to the baseline year (1998) is the marginal cost of capital (the interest rate of the lease), which is 4%. The interest rate at which the NPV equals zero is especially interesting. This is called the internal rate of return (IRR) of the project. For all interest rates below the IRR, the NPV will generate positive values. In order to determine the IRR, we construct an NPV profile for different interest rates and locate the rate where the NPV crosses the X axis where the NPV $= 0$ (Fig. 8.). It shows that the chemistry automation project can tolerate interest rates up to 18.0% (the IRR) and still generate positive returns.

The Payback Period and Average Return on Investment

Although the NPV and IRR are vastly superior indicators of project profitability because of their use of discounted cash flows, the payback period and return on investment (ROI) are still key determinant of project viability by a majority of financial managers.

The Payback Period. After 8 years the raw dollar value of positive cash flows is \$5,371,743 versus negative cash flows of \$4,0053,085. The payback period therefore:

$$8\,\text{years} \times (\$4,053,085/\$5,371,743)$$

$$8\,\text{years} \times 0.755 = 6.04\,\text{years}$$

Average ROI

Average Annual Cash Outlay:	\$4,053,085/8 years = \$506,635/ year
Average Annual Net Return:	(5,371,743 − \$4,053,085)/8 years = \$164,822
Average ROI:	(\$164,832/\$506,635) × 100 = 32.5%

CONCLUSIONS

Productivity is a key issue for labs.

The major financial benefit of automation is increased productivity.

Perianalytical automation increased our chemistry productivity by 120% (from 5,530 to 12,190 specs/tech/year).

Perianalytical automation reduced our chemistry labor cost/test by 42% (from 66¢ to 38¢/test).

Automation is a key solution for staff shortages.

Speedy implementation, speedy labor reductions and speedy revenue generation improve financial performance.

To achieve financial success, laboratorians must understand key financial principles.

ACKNOWLEDGMENTS

We thank E. Simson for practical advice on the financial perspective and M. Gannon for teaching us the meaning and calculation of the Net Present Value.

BIBLIOGRAPHY

Cited References

1. Eggert AA. Analytical methods, automated. In: Webster JG, editor. Encyclopedia of Medical Devices and Instrumentation. Hoboken (NJ): John Wiley & Sons; 1988.
2. Narayanan S. The preanalytical phase: an important component of laboratory medicine. Am J Clin Pathol 2000;113:429–452.
3. Labcorp directory of services and interpretive guide 2003.
4. Ford A. Latest chemistry wish list in low volume labs. CAP today 2004; April 44–58.
5. Lee-Lewandrowski E, Lewandrowski K. Contemporary instruments in the clinical laboratory: A brief overview. In: Lewandrowski K, editor. Clinical chemistry. Philadelphia: Lippincott Williams & Wilkins; 2002.
6. Aller R. Chemistry analyzers. CAP today 1999; July 58–83.
7. Adesoji BB, Peterson JR. Immunoassay and immunoassay analyzers: A perspective from the clinical laboratory. In: Lewandrowski K, editor. Clinical chemistry. Philadelphia: Lippincott Williams & Wilkins; 2002.
8. Goslimg JP. A decade of development in immunoassay technology. Clin Chem 1990;36:1408–1427.
9. Ehrhardt V, Assmann G, Batz O, et al. Results of the multicentre evaluation of an electrochemiluminescence immunoassay for hcg on elecsys 2010. Wein Klin Wochenschr 1998;3 (Suppl): 61–67.
10. Aller RD, Smalley D. Of all analyzers, immunoassay the trickiest. CAP today 2000;April 30–64.
11. Ford A. Automated immunoassay analyzers: the latest lineup. CAP today 2003; June 72–96.
12. Jacobs E. Acute care and stat lab testing. In: Lewandrowski K, editor. Clinical chemistry. Philadelphia: Lippincott Williams & Wilkins; 2003.

13. Ford A. Choosing cost-efficiency in low-volume labs. CAP today 2003; June 32–52.
14. Markin RS. Clinical laboratory automation. In: Henry JB, editor. Clinical diagnosis and management by laboratory methods. Philadelphia: W.B. Saunders; 2001.
15. Ford A. Laboratory automation systems and work cells. CAP today 2003; May: 35–52.
16. Sasaki M. Completed automatic clinical laboratory using a sample transportation system: the belt-line system. Jpn J Clin Pathol 1984;32:119–126.
17. NCCLS laboratory automation: specimen container/specimen carrier; proposed standard. NCCLS document auto 1 P; December 1995.
18. NCCLS laboratory automation: bar codes for specimen container identification; proposed standard. NCCLS document 2 P; April 1999.
19. NCCLS laboratory automation: communications with automated clinical laboratory systems, instruments, devices and information systems; proposed standard. NCCLS document 3 P; December 1998.
20. NCCLS laboratory automation: systems operational requirements and information elements; proposed standard. NCCLS document auto 4 P; October 1999.
21. NCCLS laboratory automation; electromechanical interface; proposed standard. NCCLS document auto 5 P; April 1999.
22. Markin RS, Whalen SA. Laboratory automation; trajectory, technology and tasks. Clin Chem 2000;46:764–771.
23. Sarkozi L, Simson E, Ramanathan L. The effects of total laboratory automation on the management of a clinical chemistry laboratory. Retrospective analysis of 36 years. Clinica Chimica Acta 2003;329:89–94.

See also BLOOD COLLECTION AND PROCESSING; COMPUTERS IN THE BIOMEDICAL LABORATORY; CYTOLOGY, AUTOMATED; DIFFERENTIAL COUNTS, AUTOMATED.

ANALYZER, OXYGEN.

See OXYGEN ANALYZERS.

ANESTHESIA MACHINES

ROBERT LOEB
University of Arizona
Tuscon, Arizona

JEFFREY FELDMAN
Children's Hospital of Philadelphia
Philadelphia, Pennsylvania

INTRODUCTION

On October 16, 1846, W. T. G. Morton gave the first successful public demonstration of inhalational anesthesia. Using a hastily devised glass reservoir to deliver diethyl ether, he anesthetized a patient before an audience at the Massachusetts General Hospital (Fig. 1). This glass reservoir thus became the first, crude, anesthesia machine. The technology of anesthesia machines has advanced immeasurably in the ensuing 150 years. Modern anesthesia machines are used to administer inhalational anesthesia safely and precisely to patients of any age, in any state of health, for any duration of time, and in a wide range of operating environments.

Figure 1. A reproduction of the Morton Inhaler, ~1850. (Image © by the Wood Library-Museum of Anesthesiology, Park Ridge, Illinois.)

The term anesthesia machine colloquially refers to all of the medical equipment used to deliver inhalational anesthesia. Inhalational anesthetics are gases that, when inhaled, produce a state of general anesthesia, a drug-induced reversible loss of consciousness during which the patient is not arousable, even in response to painful stimulation. Inhalational anesthetics are supplied as either compressed gases (e.g., nitrous oxide), or volatile liquids (e.g., diethyl ether, sevoflurane, or desflurane). In recent years, the anesthesia machine has been renamed the anesthesia delivery system, or anesthesia workstation because modern devices do more than simply deliver inhalational anesthesia. Defined precisely, the term "anesthesia machine" specifically refers to that component of the anesthesia delivery system that precisely mixes the compressed and vaporized gases that are inhaled to produce anesthesia. Other components of the anesthesia delivery system include the ventilator, breathing circuit, and waste gas scavenger system. Anesthesia workstations are anesthesia delivery systems that also incorporate patient monitoring and information management functions (Fig. 2).

The most obvious goals of general anesthesia are to render a patient unaware and insensible to pain so that surgery or other medically necessary procedures can be performed. In the process of achieving these goals, potent medications are administered that interfere with normal body functions, most notably circulation of blood and the ability to breathe (see the text box Typical Process of Delivering General Anesthesia). The most important goal of anesthesia care is therefore to keep the patient safe and free from injury.

Patient safety is a major principle guiding the design of the anesthesia workstation. Precise control of the dose of anesthetic gases and vapors reduces the risk of administering an overdose. The ventilator and breathing circuit are fundamental components of the anesthesia delivery system designed to allow for continuous delivery of oxygen to the lungs and removal of exhaled gases. To fulfill national and international standards, anesthesia delivery systems must have essential safety features and meet specified minimum performance criteria (1–6)

Typical Process of Delivering General Anesthesia

Check the anesthesia delivery system for proper function: At the start of each day, the anesthesia provider places disposable components on the breathing circuit and performs an equipment check to ensure proper function of the anesthesia workstation (7).

Identify the patient and confirm the surgical site: Healthcare institutions are required to have formal procedures to identify patients and the site of surgery before the patient is anesthetized.

Establish venous access to administer medications and fluids: Using this catheter, drugs can be administered intravenously and fluids can be given to replace loss of blood or other body fluids.

Attach physiologic monitors: Monitoring the effects of anesthesia on the body is of paramount importance to guide the dose of anesthetic given and to keep the patient safe. Typical monitors include a blood pressure cuff, electrocardiogram, and pulse oximeter. Standards require that additional monitors be used during most anesthesia care (8).

Have the patient breathe 100% oxygen through a mask and circuit attached to the anesthesia machine: A tightly fitting mask is held over the patient's face while 100% oxygen is administered using the anesthesia machine. The goal is to eliminate the nitrogen in the lungs and provide a reservoir of oxygen to sustain the patient from the time anesthesia is induced until mechanical ventilation is established.

Inject a rapidly acting sedative–hypnotic medicine into the patient's vein: This injection induces general anesthesia and often causes the patient to stop breathing. Typical induction medications (e.g., thiopental, propofol) are quickly redistributed and metabolized, so additional anesthetics must be administered shortly thereafter to maintain anesthesia.

Breathe for the patient: This is typically accomplished by holding a mask attached to the breathing circuit tightly over the patient's face and squeezing the bag on the anesthesia machine to deliver oxygen to the lungs. This process is also known as manual ventilation.

Inject a neuromuscular blocking drug to paralyze the patient's muscles: Profound muscle relaxation makes it easier for the anesthesia provider to insert a tracheal tube into the patient's trachea. Neuromuscular blockers are also often used to make it easier for the surgeon to perform the procedure.

Insert a tube into the patient's trachea: This step is called endotracheal intubation and is used to establish a secure path for delivering oxygen and inhaled anesthetics to the patient's lungs as well as eliminating carbon dioxide.

Confirm correct placement of the endotracheal tube: This step is fundamental to patient safety. Numerous methods to confirm correct placement have been described. Identifying the presence of carbon dioxide in the exhaled gas is considered the best method for confirming tube placement. Continuous monitoring of carbon dioxide in the exhaled gases is considered a standard of care during general anesthesia.

Deliver anesthetic agents: General anesthesia is typically maintained with inhaled anesthetic gases. Dials are adjusted on the anesthesia machine to dispense a specified concentration of anesthetic vapor mixed with oxygen and air or nitrous oxide.

Begin mechanical ventilation: The anesthesia delivery system is switched from spontaneous to mechanical ventilation mode, and a ventilator, built into the anesthesia delivery system, is set to breathe for the patient. This frees the anesthesia provider's hands and ensures that the patient breathes adequately during deeper levels of anesthesia and while under the effect of neuromuscular blockers. The ability to deliver anesthetic gases while providing mechanical ventilation is a unique feature of the anesthesia machine.

Adjust ventilation and depth of anesthesia: During the case, the gas flows are reduced to minimize anesthetic usage. The inhaled anesthetic concentration is adjusted to optimize the depth of anesthesia in response to changing levels of surgical stimulus. The ventilator settings are tuned to optimize the patient's ventilation and oxygenation status. Information form the physiologic monitors helps to guide these adjustments.

Establish spontaneous ventilization: Toward the end the operation, the magnitude of ventilation is decreased. The patient responds by starting to breathe spontaneously, at which time the anesthesia delivery system is switched from mechanical to spontaneous ventilation mode and the patient continues to breath from the bag on the anesthesia machine.

Remove the endotracheal tube: At the end of the case, the anesthetic gases are turned off and the patient regains consciousness. The endotracheal tube is removed and the patient breathes oxygen from a cylinder while being transported to the recovery area.

System Overview

Anesthesia delivery systems allow anesthesia providers to achieve the following goals:

1. Precisely deliver a prescribed concentration of inhaled gases to the patient.
2. Support multiple modes of ventilation (i.e., spontaneous, manually assisted, and mechanically controlled).
3. Precisely deliver a wide variety of prescribed ventilator parameters.
4. Conserve the use of anesthetic vapors and gases.
5. Minimize contamination of the operating room atmosphere by anesthetic vapors and gases.
6. Minimize the chance of operator errors.
7. Minimize patient injury in the event of operator error or equipment malfunction.

Figure 2. Four contemporary anesthesia workstations. The top two are manufactured by GE Healthcare, and the bottom two by Draeger Medical.

These goals will be discussed further in the following section, which describes the major components of the anesthesia delivery system. The following overview of anesthesia delivery system function will refer to these goals.

The anesthesia delivery system consists of four components: a breathing circuit, an anesthesia machine, a waste gas scavenger system, and an anesthesia ventilator. The breathing circuit is the functional center of the system, since it is physically and functionally connected to each of the other components and to the patient's airway (Fig. 3). There is a one-way flow of gas from the anesthesia machine into the breathing circuit, and from the breathing circuit into the scavenger system. There is a bidirectional flow of gas between the breathing circuit and the patient's lungs, and between the breathing circuit and the anesthesia ventilator or reservoir bag. The ventilator and the reservoir bag are functionally interchangeable units, which are used during different modes of ventilation (Goal 2). During spontaneous and manually assisted modes of ventilation, the elastic reservoir bag is used as a source of inspired gas and a low impedance reservoir for exhaled gas. The anesthesia ventilator is used during mechanically controlled ventilation to automatically inflate the lungs using prescribed parameters (Goal 3).

During inhalation, gas flows from the anesthesia ventilator or reservoir bag through the breathing circuit to the patient's lungs. The patient's bloodstream takes up a small portion of gas (e.g., oxygen and anesthetic agent) from the lungs and releases carbon dioxide (CO_2) into the lungs.

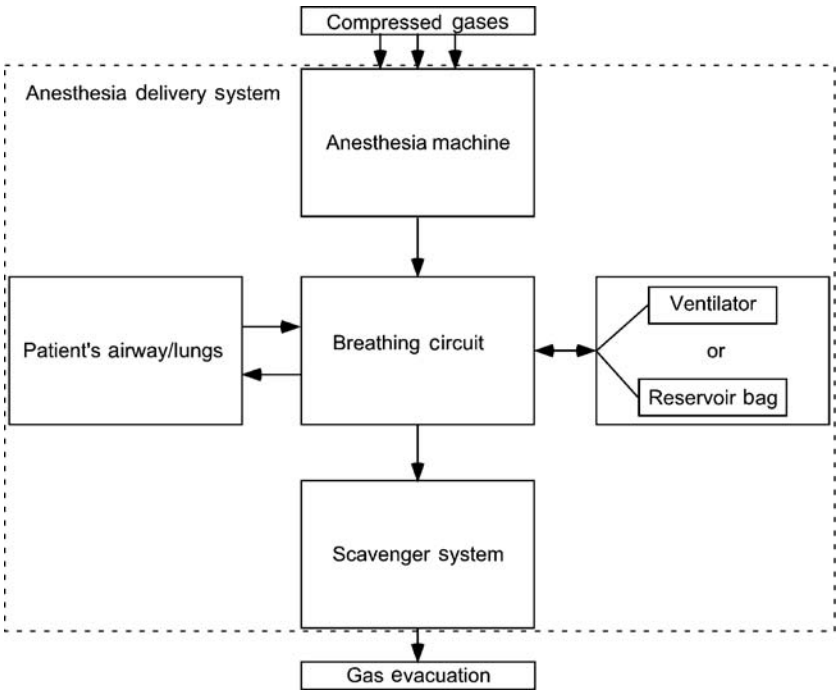


Figure 3. Block diagram of anesthesia delivery system components. The arrows show the direction of gas flow between components.

During exhalation, gas flows from the patient's lungs through the breathing circuit back to the anesthesia ventilator or reservoir bag. This bulk flow of gas, between the patient and the ventilator or reservoir bag, constitutes the patient's pulmonary ventilation; the volume of each breath is referred to as tidal volume, and the total volume exchanged during one minute is referred to as minute volume.

Over time, the patient absorbs oxygen and anesthetic agents from, and releases CO_2 to, the gas in the breathing circuit. Without intervention, the gas within the breathing circuit would progressively decrease in total volume, oxygen concentration, and anesthetic concentration. The anesthesia provider, therefore, dispenses fresh gas into the breathing circuit, replacing the gas absorbed by the patient. Using the anesthesia machine, the anesthesia provider precisely controls both the flow rate and the concentration of various gases in the fresh gas (Goal 1). The anesthesia machine is capable of delivering a total fresh gas flow that far exceeds the volume of gas absorbed by the patient. When higher fresh gas flows are used (for example, to rapidly change the concentration of gases in the breathing circuit), the excess gas is vented into the scavenger system to be evacuated from the operating room (Goal 5).

To conserve the use of anesthetic gases (Goal 4), the anesthesia provider will use a fresh gas flow rate that is significantly lower than the patient's minute volume. In this situation, the patient reinhales gas that they had previously exhaled into the breathing circuit (this is called rebreathing). Carbon dioxide absorbent contained within the breathing circuit prevents the patient from rebreathing CO_2, which would be deleterious. All other gases (oxygen, nitrous oxide, nitrogen, and anesthetic vapors) can be rebreathed safely.

During the course of a typical anesthetic, the anesthesia provider will use a relatively high fresh gas flow at the beginning and end of the anesthetic when a rapid change in anesthetic concentration is desired, and a lower fresh gas flow when little change in concentration is desired. The technique of closed circuit anesthesia refers to the process of adjusting the fresh gas flow to exactly match the amount of gas used by the patient so that no gas is vented to the scavenging system.

Because anesthesia delivery systems provide critical life support functions to unconscious patients, equipment malfunctions and user errors can have catastrophic consequences. In 1974, the American National Standards Institute published an anesthesia machine standard that specified minimum performance and safety requirements for anesthesia gas machines (Goals 6 and 7). That standard was a landmark one, in that it was the first systematic approach to standardize the safety requirements for a medical device. Similar standards have since been written for other medical equipment, and the anesthesia machine standards have been regularly updated.

Breathing Circuit (Semiclosed Circle System)

The semiclosed circle system is the most commonly used anesthesia breathing circuit, and the only type that will be discussed in this article. It is so named because expired gases can be returned to the patient in a circular fashion (Fig. 4). The components of the circle system include a carbon dioxide absorber canister, two one-way valves, a reservoir bag, an adjustable pressure-limiting valve, and tubes that connect to the patient, ventilator, anesthesia machine, and scavenger system.

During inspiration, the peak flow of gas exceeds 25 $L \cdot min^{-1}$, far in excess of the rate of fresh gas supply. As a result, the patient will inspire both fresh gas and gas stored in the reservoir bag or ventilator bellows. Inspired gas travels through the carbon dioxide absorber canister, past the one-way inspiratory valve, to the patient. During

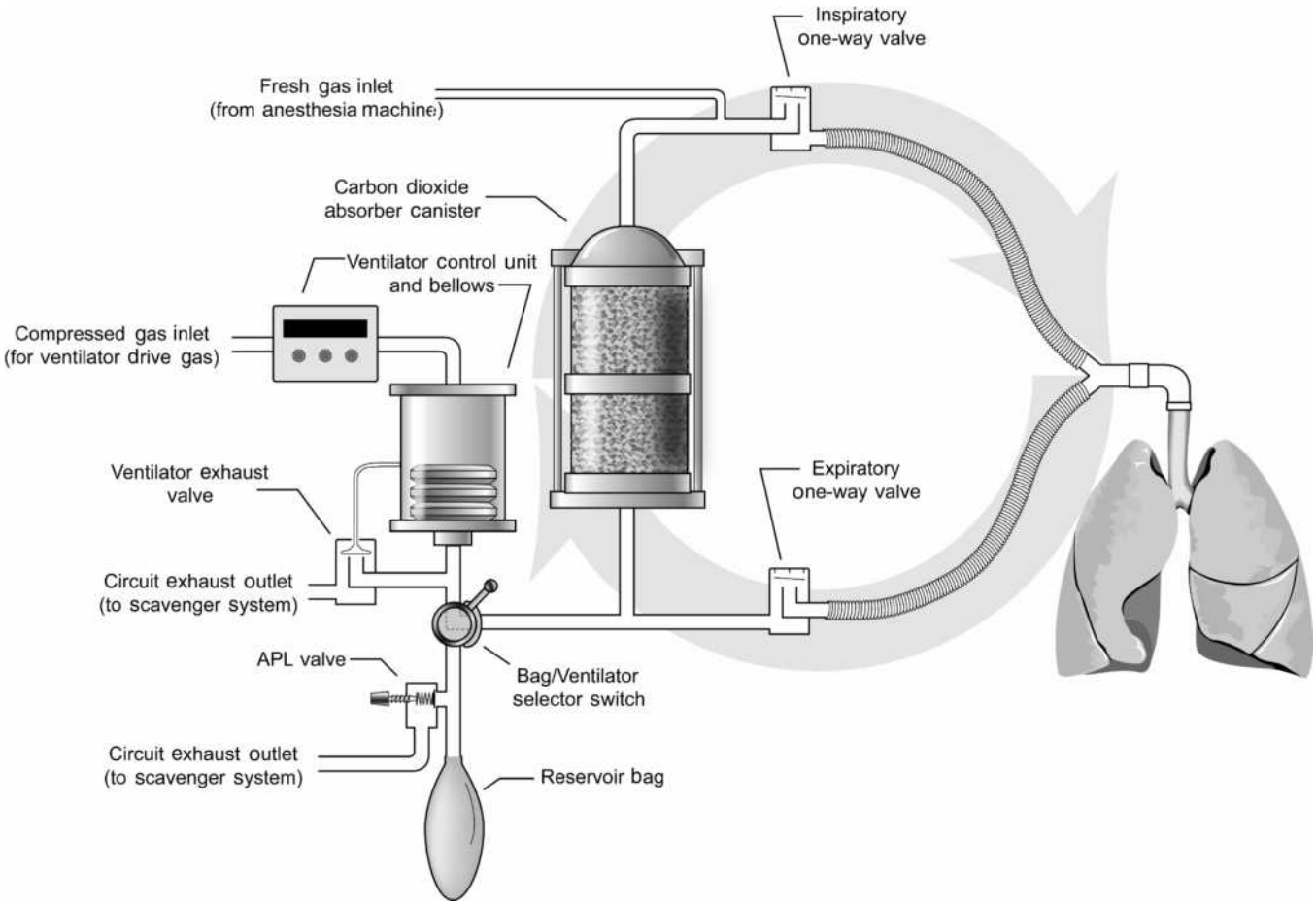


Figure 4. This schematic of the circle breathing circuit shows the circular arrangement of components. The one-way valves permit flow in only one direction.

exhalation, gas travels from the patient, past the one-way expiratory valve, to the reservoir bag (or ventilator bellows, depending upon the position of the bag–ventilator selector switch). The one-way valves establish the direction of gas flow in the breathing circuit. Carbon dioxide is not rebreathed because exhaled gas is directed through the carbon dioxide absorber canister prior to being reinhaled. Fresh gas from the anesthesia machine flows continuously into the breathing circuit. During inhalation, this gas joins with the inspiratory flow and is directed toward the patient. During exhalation, the fresh gas enters the breathing circuit and travels retrograde through the carbon dioxide absorber canister toward the reservoir bag (it does not travel toward the patient because the inspiratory one-way valve is closed during exhalation). Thus, during exhalation, gas enters the reservoir bag from the expiratory limb and from the carbon dioxide absorber canister. Once the reservoir bag is full, excess returning gas is vented out the adjustable pressure-limiting (APL) valve to the scavenger system (when the ventilator is used, the excess gas is vented out the ventilator exhaust valve). The total fresh gas flow will therefore control the amount of gas that is rebreathed. At high fresh gas flows, the exhaled gases are washed out through the scavenging system between each inspiration. At low fresh gas flow, very little exhaled gas is forced out to the scavenging system and most of the exhaled gas is reinhaled in subsequent breaths.

CIRCLE SYSTEM COMPONENTS

CO_2 Absorbents

Alkaline hydroxides of sodium, potassium, calcium, and barium in varying concentrations are most commonly used as carbon dioxide absorbents. These alkaline hydroxides irreversibly react with carbon dioxide to eventually form carbonates, releasing water and heat. Absorbent granules are 4- to 8-mesh in size (25–35 granules cm^{-3}) to maximize the surface area available for chemical reaction and minimize the resistance to gas flow through the absorber canister. Ethyl violet is incorporated into the granules as a pH indicator; fresh granules are white, while a purple color indicates that the absorbent needs to be replaced. Absorber canisters are constructed with transparent sides so that absorbent color can be easily monitored during use. Canisters have a typical capacity of 900–1200 cm^3 and the absorbent is good for ~10–30 h of use, depending on the operating conditions.

Many of the absorbent materials have the potential to interact with anesthetic agents to degrade the anesthetics

and produce small amounts of potentially toxic gases, such as carbon monoxide. This is especially true if the absorbents are allowed to dessicate by exposure to high flows of dry gas (e.g., leaving the fresh gas flowing on the anesthesia machine over a weekend). Periodic replacement of absorbent, especially at the end of a weekend is therefore desirable. Newer absorbent materials, which are more costly, are designed to reduce or eliminate the potential for producing toxic gases by eliminating the hydroxides of sodium, barium, and potassium.

Unidirectional Valves

The inspiratory and expiratory one-way valves are simple, passive devices. Each has an inlet tube that is capped by a valve disk. When the pressure in the inlet tube exceeds that in the outlet tube, the valve opens to allow gas to flow downstream. The valve disks are light in weight to minimize gas flow resistance. Each valve has a clear dome to allow visual monitoring of valve function. Rarely, valves malfunction by failing to open or close properly. Carbon dioxide rebreathing can occur if either valve becomes incompetent (i.e., fails to close properly). This can occur if a valve disk becomes warped, sticks open due to humidity, or fails to seat properly.

Reservoir Bag

The reservoir bag is an elastic bag that serves three functions in the breathing circuit. First, it is a compliant element of an otherwise rigid breathing circuit that allows changes in breathing circuit gas volume without changes in breathing circuit pressure. Second, it provides a means for manually pressurizing the circuit to control or assist ventilation. Third, it provides a safety limit on the peak pressure that can be achieved in the breathing circuit. It acts as a pressure-limiting device in the event that fresh gas inflow exceeds APL valve outflow. Reservoir bags are designed such that, at fresh gas flow rates below 15 $L \cdot min^{-1}$, the breathing circuit pressure will remain < 35 cm H_2O (3.4 kPa) until the bag reaches more than twice its full capacity. Yet, inspiratory pressures up to 70 cm H_2O (6.9 kPa) can be achieved by quickly compressing the reservoir bag.

APL Valve

The APL valve (euphemistically referred to as the pop-off valve) is a spring-loaded device that controls the flow of gas from the breathing circuit to the scavenger system. The valve opens when the pressure gradient from the circuit to the scavenger exceeds the force exerted by the spring (as discussed later, the pressure in the scavenger system is regulated to be equal to atmospheric pressure plus or minus a few cm H_2O). When the patient is breathing spontaneously, the anesthesia practitioner minimizes the spring tension allowing the valve to open with minimal end-expiratory pressure (typically < 3 cm H_2O, or 0.3 kPa). When the anesthesia practitioner squeezes the reservoir bag to manually control or assist ventilation, the APL valve opens during inhalation. Part of the gas exiting the reservoir bag escapes to the scavenger system and the remainder is directed toward the patient. By turning a knob, the anesthesia practitioner increases the pressure on the spring so that the APL valve remains closed until the pressure in the circuit achieves a level that is adequate to inflate the patients lungs; the APL valve thus opens toward the end of inhalation, once the lungs are adequately inflated. Continual adjustment of the APL valve is sometimes needed to adapt to changing fresh gas flow rate, circuit leaks, pulmonary mechanics, and ventilation parameters.

Bag–Ventilator Selector Switch

During mechanical ventilation, the reservoir bag and APL valve are disconnected from the breathing circuit and an anesthesia ventilator is connected to the same spot. Modern breathing circuits have a selector switch that quickly toggles the connection to either the ventilator or the reservoir bag and APL valve.

VIRTUES AND LIMITATIONS OF THE CIRCLE BREATHING CIRCUIT

Primary advantages of the circle breathing system over other breathing circuits include conservation of anesthetic gases and vapors, ease of use, and humidification and heating of inspired gases.

As stated previously, anesthetic agents are conserved when very low fresh gas flows are used with the circle breathing system. The minimum adequate flow is one that just replaces the gases taken up by the patient; for a normal adult, flows below 0.5 $L \cdot min^{-1}$ can be achieved during anesthesia maintenance. It is customary to use higher fresh gas flow rates in the range of 1–2 $L \cdot min^{-1}$, but this is still well below typical minute ventilation rates of 5–10 $L \cdot min^{-1}$ which is the fresh gas flow that would be required for a nonrebreathing ventilation system.

The circle breathing circuit is easy to use because the same fresh gas settings can be used with patients of various sizes. A 100 kg adult and a 1 kg infant can each be anesthetized with a circle breathing system and a fresh gas maintenance flow rate of 1–2 $L \cdot min^{-1}$. Since the larger patient would take up more anesthetic agent and more oxygen, and would give off more carbon dioxide, higher *minimal* flows would be required for the larger patient and the carbon dioxide absorbent would become exhausted quicker. Also, for convenience, a smaller reservoir bag and smaller bore breathing tubes would be selected for the smaller patient. But, otherwise, the system would function similarly for both patients.

Humidification and warming of inspired gases is another advantage of rebreathing. Fresh gas is mixed from compressed gases that contain zero water vapor, and breathing this dry gas can have detrimental effects on lung function. But, within the circle breathing system, inspired gas is humidified by the admixture of rebreathed gas, and by the water vapor that forms as a byproduct of carbon dioxide absorption. Both of these mechanisms also act to warm the inspired gas. By using low flows, enough heat and humidity is conserved to eliminate the need to actively heat and humidify inspired gas.

Most disadvantages of the circle breathing system are due to the large circuit volume. Internal volumes are primarily

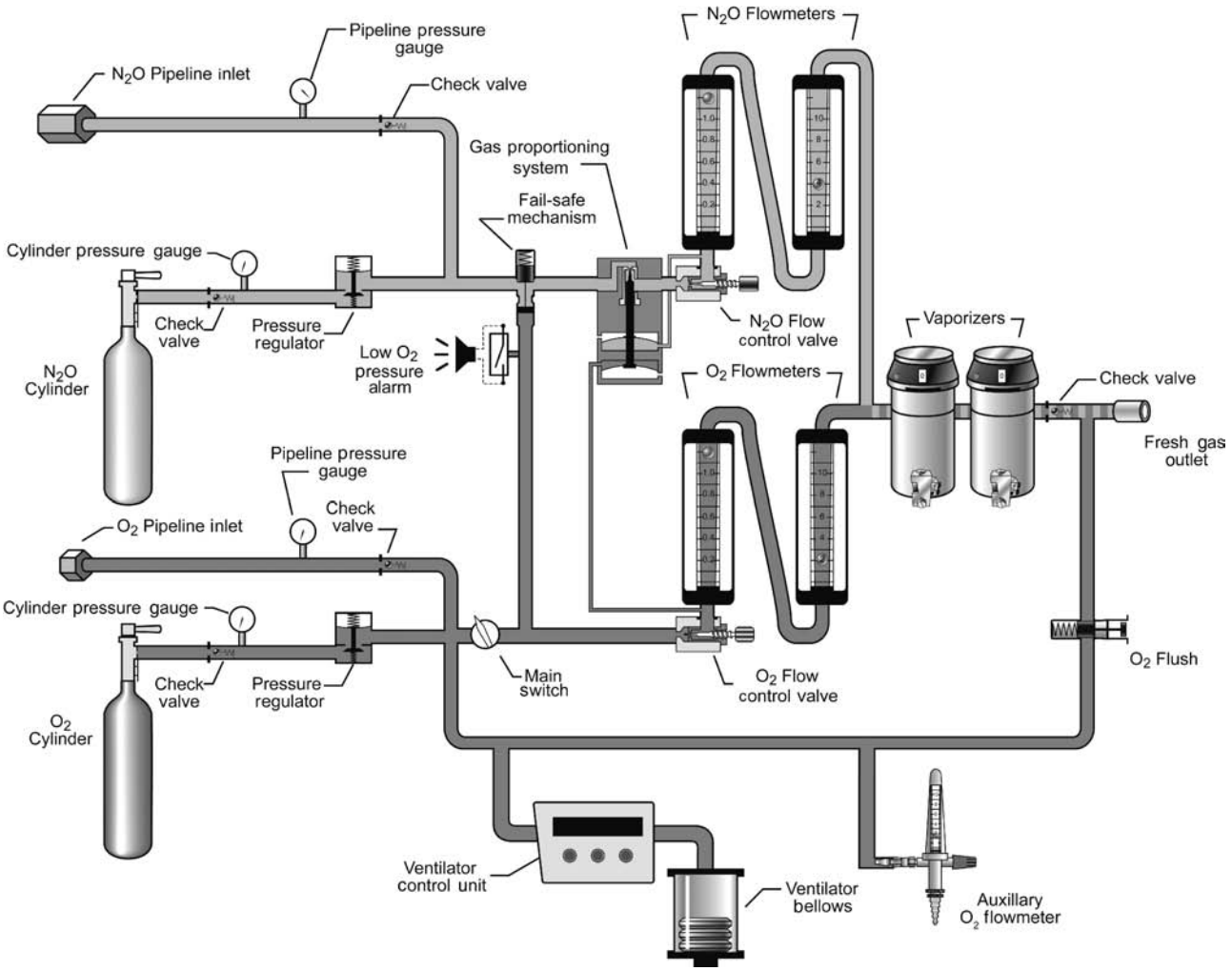


Figure 5. Schematic showing the internal piping and placement of components within the anesthesia machine. Dark gray indicates oxygen (O_2) and light gray indicates nitrous oxide (N_2O).

determined by the sizes of the absorbent canister, reservoir bag, and breathing hoses; 3–6 L are typical. Large circuits are physically bulky. They also increase the time required to change inspired gas concentrations because the large reservoir of previously exhaled gas is continually added to fresh gas. Finally, large circuits are more compliant, which degrades the efficiency and accuracy of ventilation. This effect will be discussed further in the section on ventilators.

Anesthesia Machine

The anesthesia machine is used to accurately deliver into the breathing circuit a precise flow and concentration of gases and vapors. Anesthesia machines are manufactured to deliver various compressed gases; all deliver oxygen, most deliver nitrous oxide or air, some deliver helium or carbon dioxide. They have one or more vaporizers that convert liquid anesthetic agents into anesthetic vapors; currently used inhaled vapors include halothane, enflurane, isoflurane, sevoflurane, and desflurane. Anesthesia machines include numerous safety features that alert the anesthesia provider to malfunctions and avert use errors.

The anesthesia machine is a precision gas mixer (Fig. 5). Compressed gases enter the machine from the hospital's centralized pipeline supply or from compressed gas cylinders. The compressed gases are regulated to specified pressures, and each passes through its own flow controller and flow meter assembly. The compressed gases then are mixed together and may flow through a single vaporizer where anesthetic vapor is added. The final gas mixture then exits the common gas outlet (also called the fresh gas outlet) to enter the breathing circuit.

ANESTHESIA MACHINE COMPONENTS

Compressed Gas Inlets

Compressed gases from the hospital pipeline system or from large compressed gas cylinders enter the anesthesia machine through flexible hoses. The inlet connector for each gas is unique in shape to prevent the connection of the wrong supply hose to a given inlet. The standardized design of each hose-inlet connector pair conforms to the Diameter Indexed Safety System (DISS) (9).

Anesthesia machines also have inlet yokes that hold small compressed gas cylinders; these cylinders provide compressed gas for emergency backup and for use in locations without piped gases. Each yoke is designed to prevent incorrect placement of a cylinder containing another gas. Two pins located in the yoke must insert into corresponding holes on the cylinder valve stem. The standardized placement of these pins and corresponding holes, referred to as the Pin Indexed Safety System (PISS), is unique for each gas (10).

Pressure Regulators And Gauges

Gauges on the front panel of the anesthesia machine display the cylinder and pipeline inlet pressures of each gas. Gases from the pipeline inlets enter the anesthesia machine at pressures of 45–55 psig (310–380 kPa), whereas gases from the compressed gas cylinders enter at pressures up to 2000 psig (1379 kPa). (Pressure conversion factors: 1 psig = 0.068 atm = 51.7 mmHg = 70.3 cm H_2O = 6.89 kPa.) Pressure regulators on each cylinder gas inlet line reduce the pressure from each cylinder to $\sim$ 45 psig (310 kPa). The pressure regulators provide a relatively constant outlet pressure in the presence of a variable inlet pressure, which is important since the pressure within a gas cylinder declines during use. Lines from the pipeline inlet and the cylinder inlet (downstream of the pressure regulator) join to form a common source line for each gas. Gases are preferentially used from the pipelines, since the pressure regulators are set to outlet pressures that are less than the usual pipeline pressures.

Flow Controllers And Meters

A separate needle-valve controls the flow rate of each compressed gas. Turning a knob on the front panel of the anesthesia machine counterclockwise opens the needle valve and increases the flow; turning it clockwise decreases or stops the flow. A flowmeter assembly, located above each flow-control knob, shows the resulting flow rate. The flowmeter consists of a tapered glass tube containing a movable float; the internal diameter of the tube is larger at the top than at the bottom. Gas flows up through the tube, which is vertically aligned, and in doing so blows the float higher in the tube. The float balances in midair partway up the tube when its weight equals the force of the gas traveling through the space between the float and the tube. Thus, the height to which the float rises within the tube is proportional to the flow rate of the gas. Flow rate is indicated by calibrated markings on the tube alongside the level of the float.

Each flowmeter assembly is calibrated for a specific gas. The density and viscosity of the gas significantly affects the force generated in traveling through the variable-sized annular orifice created by the outer edge of the float and the inner surface of the tube. Temperature and barometric pressure affect gas density, and major changes in either can alter flowmeter accuracy. Accuracy is also impaired by dirt or grease within the tube, static electricity between the float and the tube, and nonvertical alignment of the tube.

To increase precision and accuracy, some machines indicate gas flow rate past a single needle valve using two flowmeter assemblies, one for high flows and the other for low flows. These flowmeters are connected in series and the flow rate is indicated on one flowmeter or the other. A flow rate below the range of the high-flow meter shows an accurate flow rate on the low flow meter and an unreadable low flow rate on the high flow meter. While, a flow rate that exceeds the range of the low-flow meter shows an accurate flow rate on the high flow meter and an unreadable high flow rate on the low flow meter.

Each gas, having passed through its individual flow controller and meter assembly, passes into a common manifold before continuing on. Only the individual gas flow rates are indicated on the flowmeters; the user must calculate the total gas flow rate and the percent concentration of each gas in the mixture.

Vaporizers

Vaporizers are designed to add an accurate amount of volatilized anesthetic to the compressed gas mixture. Anesthetic vapors are pharmacologically potent, so low concentrations (generally $< 5\%$) are typically needed. The volatilized gases contribute to the total gas flow rate and dilute the concentration of the other compressed gases. The user can calculate these effects since they are not displayed on the machine front panel; luckily, these are generally negligible and can be ignored. Even though most anesthesia machines have multiple vaporizers, only one is used at a time; interlock mechanisms prevent a vaporizer from being turned on when another vaporizer is in use. Vaporizers are anesthetic agent specific and keyed filling systems prevent filling a vaporizer with the wrong liquid anesthetic.

All current anesthesia machines have direct-setting vaporizers that add a specified concentration of a single anesthetic vapor to the compressed gas mixture. Variable-bypass vaporizers are the most common (Fig. 6). In these, the inflowing compressed gas mixture is split into two streams. One stream is directed through a bypass channel and the other is directed into a chamber within the vaporizer that contains liquid anesthetic agent. The gas entering the vaporizing chamber becomes saturated with anesthetic vapor at a concentration that depends on the vapor pressure of the particular liquid anesthetic. For example, sevoflurane has a vapor pressure of 157 mmHg (20.9 kPa) at 20 °C, so the gas within the vaporizing chamber is about 20% sevoflurane (at sea level). This highly concentrated anesthetic mixture exits the chamber (now, at a flow rate greater than that entering the chamber, due to the addition of anesthetic vapor) to join, and be diluted by, gas that traversed the bypass channel. A dial on the vaporizer controls the delivered anesthetic concentration by regulating the resistance to flow along each path. For example, setting a sevoflurane vaporizer to a dialed concentration of 1% splits the inflowing compressed gas mixture so that one-twenty-fourth of the total is directed through the vaporizing chamber and the remainder is directed through the bypass. Direct-reading variable-bypass vaporizers are calibrated for a specific agent, since each anesthetic liquid has a different vapor pressure. Vapor pressure varies with

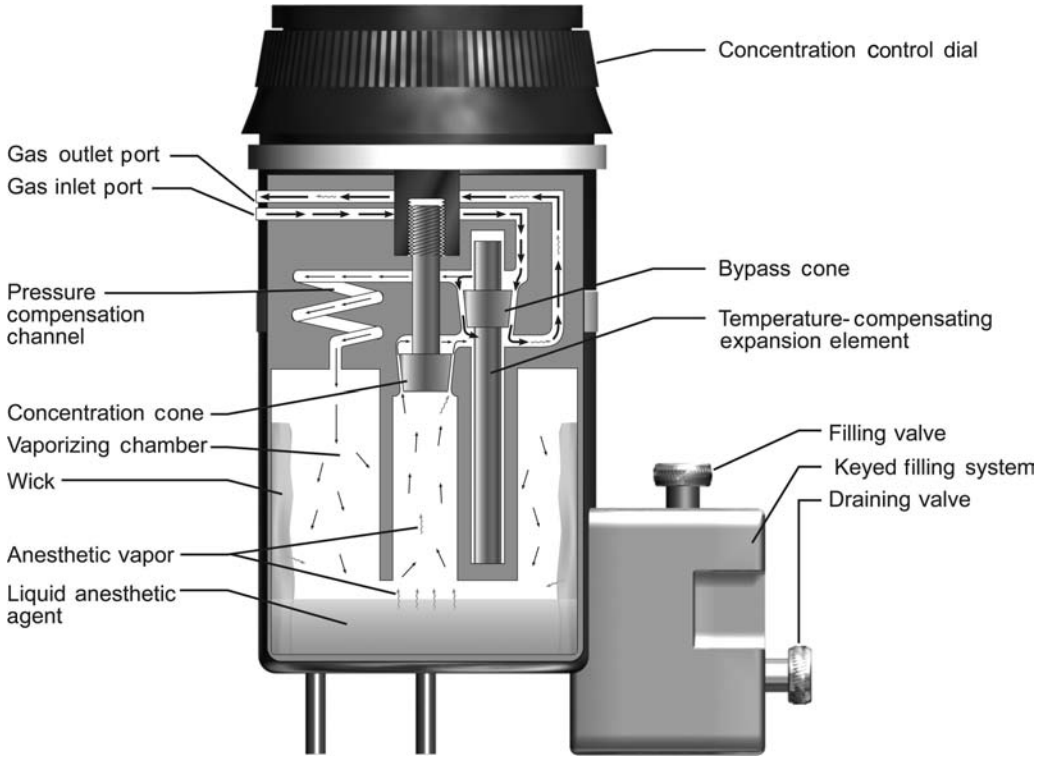


Figure 6. Schematic of a variable-bypass vaporizer. Arrows indicate direction of gas flow; heavier arrows indicate larger flow rates. Gas enters the Inlet Port and is split at the Bypass Cone into two streams. One stream is directed through a bypass channel and the rest enters the Vaporizing Chamber. Gas entering the Vaporizing Chamber equilibrates with Liquid Anesthetic Agent to become saturated with Anesthetic Vapor. This concentrated anesthetic mixture exits the chamber to join, and be diluted by, gas that traversed the bypass channel. The Concentration Control Dial is attached to the Concentration Cone, which regulates resistance to flow exiting the Vaporizing Chamber and thus controls the anesthetic concentration dispensed from the Outlet Port.

temperature, so vaporizers are temperature compensated; at higher temperatures, a temperature sensitive valve diverts more gas through the bypass channel. Vaporizers are designed to ensure that the gas within the liquid-containing chamber is saturated with anesthetic vapor. A cotton wick within the chamber promotes saturation by increasing the surface area of the liquid. Thermal energy is required for liquid vaporization (heat of vaporization). To minimize cooling of the anesthetic liquid, vaporizers are constructed of metals with high specific heat and high thermal conductivity so that heat is transferred easily from the surroundings. The output of variable-bypass vaporizers varies with barometric pressure; delivered concentration increases as barometric pressure decreases.

Desflurane vaporizers are designed differently because desflurane has such a high vapor pressure (664 mmHg, or 88.5 kPa, at 20 °C) and low boiling point (22.8 °C). Uncontrollably high output concentrations could easily occur if desflurane were administered at room temperature from a variable-bypass vaporizer. In a desflurane vaporizer, the liquid desflurane is electrically heated to a controlled temperature of 39 °C within a pressure-tight chamber. At this temperature, the vapor pressure of desflurane is 1500 mmHg (200 kPa) and the anesthetic vapor above the liquid is a compressed gas. The concentration dial on the vaporizer regulates a computer-assisted flow proportioning mechanism that meters pressurized desflurane into the incoming gas mixture to achieve a set output concentration of desflurane vapor. Room temperature does not affect the output concentration of the vaporizer, nor does barometric pressure. The vaporizer requires electrical power for the heater, the onboard computer, and two electronic valves.

Safety Systems

By written standard, the anesthesia machine has numerous safety systems designed to prevent use errors. Some of these, such as the DISS and PISS systems to prevent compressed gas misconnections, interlock mechanisms to prevent simultaneous use of multiple vaporizers, and keyed filler systems to prevent misfilling of vaporizers, have already been discussed. Others are presented, below.

Failsafe Mechanism And Oxygen Alarm. The anesthesia machine has a couple of safety systems that alert the user and stop the flow of other gases when the oxygen supply runs out (for example, when an oxygen tank becomes depleted). An auditory alarm sounds and a visual message appears to alert the user when the oxygen supply pressure falls below a predetermined threshold pressure of ~ 30 psig (207 kPa). A failsafe valve in the gas line supplying each flow controller-meter assembly, except oxygen, stops the

flow of other gases. The failsafe valve is either an on–off valve or a pressure-reducing valve that is controlled by the pressure within the oxygen line. When the oxygen supply pressure falls below the threshold level, the failsafe valves close to stop the flow, or proportionally reduce the supply pressure, of all the other gases. This prevents administration of hypoxic gases (e.g., nitrous oxide, helium, nitrogen, carbon dioxide) without oxygen, which could rapidly cause injury to the patient, but it also prevents administration of air without oxygen. The failsafe mechanisms do *not* prevent delivery of hypoxic gas mixtures in the presence of adequate oxygen supply pressure; the gas proportioning system, described below, prevents this.

Gas Proportioning System. Anesthesia machines are equipped with proportioning systems that prevent the delivery of high concentrations of nitrous oxide, the most commonly used non-oxygen containing gas. A mechanical or pneumatic link between the oxygen and nitrous oxide lines ensures that nitrous oxide does not flow without an adequate flow of oxygen. One such mechanism, the Datex-Ohmeda Link-25 system, is a chain linkage between sprockets on the nitrous oxide and oxygen flow needle valves. The linkage is engaged whenever the nitrous oxide is set to exceed three-times the oxygen flow, or when the oxygen flow is set to less than one-third of the nitrous oxide flow; this limits the nitrous oxide concentration to a maximum of 75% in oxygen. Another mechanism, the Draeger Oxygen Ratio Monitor Controller (ORMC), is a slave flow control valve on the nitrous oxide line that is pneumatically linked to the oxygen line. This system limits the flow of nitrous oxide to a maximum concentration of $72 \pm 3\%$ in oxygen. Both of the above systems control the ratios of nitrous oxide and oxygen, but do not compensate for other gases in the final mixture; a hypoxic mixture (oxygen concentration $< 21\%$) could be dispensed, therefore, if a third gas were added in significant concentrations.

Oxygen Flush. Each anesthesia machine has an oxygen flush system that can rapidly deliver 45–70 $L \cdot min^{-1}$ of oxygen to the common gas outlet. The user presses the oxygen flush valve in situations where high flow oxygen is needed to flush anesthetic agents out of the breathing circuit, rapidly increase the inhaled oxygen concentration, or compensate for a large breathing circuit leak (for example, during positive pressure ventilation of the patient with a poorly fitted face mask). The oxygen flush system also serves as a safety system because it bypasses most of the internal plumbing of the anesthesia machine (e.g., safety control valves, flow controller-meter assemblies, and vaporizers) and because it is always operational, even when the anesthesia machine's master power switch is off.

Monitors and User-Interface Features. Written standards specify that all anesthesia machines must be equipped with essential safety monitors and user-interface features. To protect against hypoxia, each has an integrated oxygen analyzer that monitors the oxygen concentration in the breathing circuit whenever the anesthesia machine is powered on. The oxygen monitor must have an audible alarm that sounds whenever the oxygen concentration falls below a preset threshold, which cannot be set $< 18\%$. To protect against dangerously high and low airway pressures, the breathing circuit pressure is continuously monitored by an integrated system that alarms in the event of sub-atmospheric airway pressure, sustained high airway pressure, or extremely high airway pressure. To protect against ventilator failure and breathing circuit disconnections, the breathing circuit pressure is monitored to ensure that adequate positive pressure is generated at least a few times a minute whenever the ventilator is powered on; a low airway pressure alarm (AKA disconnect alarm) is activated whenever the breathing circuit pressure does not reach a user-set threshold level over a 15 s interval. User-interface features protect against mistakes in gas flow settings. Oxygen controls are always positioned to the right of other gas flow controls. The oxygen flow control knob has a unique size and shape that is different from the other gas control knobs. The flow control knobs are protected against their being bumped to prevent accidental changes in gas flow rates. All gas flow knobs and vaporizer controls uniformly increase their settings when turned in a clockwise direction.

LIMITATIONS

Anesthesia machines are generally reliable and problem-free. Limitations include that they require a source of compressed gases, are heavy and bulky, are calibrated to be accurate at sea level, and are designed to function in an upright position within a gravitational field. Machine malfunctions are usually a result of misconnections or disconnections of internal components during servicing or transportation. Aside from interlock mechanisms that decrease the likelihood of wrong gas or wrong anesthetic agent problems, there are no integrated monitors to ensure that the vaporizers are filled with the correct agents and the flow meters are dispensing the correct gases. Likewise, except for oxygen, the gas supply pressures and anesthetic agent levels are not automatically monitored. Thus, problems can still result when the anesthesia provider fails to diagnose a problem with the compressed gas or liquid anesthetic supplies.

Ventilator

General anesthesia impairs breathing by two mechanisms, it decreases the impetus to breath (central respiratory depression), and it leads to upper airway obstruction. Additionally, neuromuscular blockers, which are often administered during general anesthesia, paralyze the muscles of respiration. For these reasons, breathing may be supported or controlled during anesthesia to ensure adequate minute ventilation. The anesthesia provider can create intermittent positive pressure in the breathing circuit by rhythmically squeezing the reservoir bag. Ventilatory support is often provided in this way for short periods of time, especially during the induction of anesthesia. During mechanical ventilation, a selector switch is toggled to disconnect the reservoir bag and APL valve from the breathing circuit and connect an anesthesia ventilator instead. Anesthesia

ventilators provide a means to mechanically control ventilation, delivering consistent respiratory support for extended periods of time and freeing the anesthesia provider's hands and attention for other tasks. Most surgical patients have normal pulmonary mechanics and can be adequately ventilated with an unsophisticated ventilator designed for ease of use. But, high performance anesthesia ventilators allow safe and effective ventilation of a wide variety of patients, including neonates and the critically ill.

Most anesthesia ventilators are pneumatically powered, electronically controlled, and time cycled. All can be set to deliver a constant tidal volume at a constant rate (volume control). Many can also be set to deliver a constant inspiratory pressure at a constant rate (pressure control). All anesthesia ventilators allow spontaneous patient breaths between ventilator breaths (intermittent mandatory ventilation, IMV), and all can provide PEEP during positive pressure ventilation (note that in some older systems PEEP is set using a PEEP-valve integrated into the expiratory limb of the breathing circuit, and is not actively controlled by the ventilator). In general, anesthesia ventilators do not sense patient effort, and thus do not provide synchronized modes of ventilation, pressure support, or continuous positive airway pressure (CPAP).

As explained above, the anesthesia delivery system conserves anesthetic gases by having the patient rebreathe previously exhaled gas. Unlike intensive care ventilators, which deliver new gas to the patient during every breath, anesthesia ventilators function as a component of the anesthesia delivery system and maintain rebreathing during mechanical ventilation. In most anesthesia ventilators, this is achieved by incorporating a bellows assembly (see Fig. 4). The bellows assembly consists of a distensible bellows that is housed in a clear rigid chamber. The bellows is functionally equivalent to the reservoir bag; it is attached to, and filled with gas from, the breathing circuit. During inspiration, the ventilator injects drive gas into the rigid chamber; this squeezes the bellows and directs gas from the bellows to the patient via the inspiratory limb of the breathing circuit. The drive gas, usually oxygen or air, remains outside of the bellows and never enters the breathing circuit. During exhalation, the drive gas within the rigid chamber is vented to the atmosphere, and the patient exhales into the bellows through the expiratory limb of the breathing circuit.

The bellows assembly also contains an exhaust valve that vents gas from the breathing circuit to the scavenger system. This ventilator exhaust valve serves the same function during mechanical ventilation that the APL valve serves during manual or spontaneous ventilation. However, unlike the APL valve, it is held closed during inspiration to ensure that the set tidal volume dispensed from the ventilator bellows is delivered to the patient. Excess gas then escapes from the breathing circuit through this valve during exhalation.

The tidal volume set on an anesthesia ventilator is not accurately delivered to the patient; it is augmented by fresh gas flow from the anesthesia machine, and reduced due to compression-loss within the breathing circuit. Fresh gas, flowing into the breathing circuit from the anesthesia machine, augments the tidal volume delivered from the ventilator because the ventilator exhaust valve, which is the only route for gas to escape from the breathing circuit, is held closed during inspiration. For example, at a fresh gas flow rate of 3 $L \cdot min^{-1}$ (50 $mL \cdot s^{-1}$), and ventilator settings of 10 breaths min^{-1} and an I/E ratio of 1:2 (inspiratory time = 2 s), the delivered tidal volume is augmented by 100 mL per breath (2 s per breath × 50 $mL \cdot s^{-1}$). Conversely, the delivered tidal volume is reduced due to compression loss within the breathing circuit. The magnitude of this loss depends on the compliance of the breathing circuit and the peak airway pressure. Circle breathing circuits typically have a compliance of 7–9 $mL \cdot cm^{-1}$ H_2O (70–90 $mL \cdot kPa^{-1}$), which is significantly higher than the typical 1–3 $mL \cdot cm^{-1}$ H_2O (10–30 $mL \cdot kPa^{-1}$) circuit compliance of intensive care ventilators, because of their large internal volume. For example, when ventilating a patient with a peak airway pressure of 20 cm H_2O (2 kPa) using an anesthesia ventilator with a breathing circuit compliance of 10 mL·cm H_2O, delivered tidal volume is reduced by 200 mL per breath.

LIMITATIONS

Until recently, anesthesia ventilators were simple devices designed to deliver breathing circuit gas in volume control mode. The few controls consisted of a power switch, and dials to set respiratory rate, inspiratory/expiratory (I/E) ratio, and tidal volume. While simple to operate, these ventilators had a number of limitations. As discussed above, delivered tidal volume was altered by peak airway pressure and fresh gas flow rate. Tidal volume augmentation was particularly hazardous with small patients, such as premature infants and neonates, since increasing the gas flow on the anesthesia machine could unintentionally generate dangerously high tidal volumes and airway pressures. Tidal volume reduction was particularly hazardous since dramatically lower than set tidal volumes could be delivered, unbeknown to the provider, to patients requiring high ventilating pressures (e.g., those with severe airway disease or respiratory distress syndrome). Worse yet, the pneumatic drive capabilities of these ventilators were sometimes insufficient to compensate for tidal volume losses due to compression within the breathing circuit; anesthesia ventilators were unable to adequately ventilate patients with high airway pressures (> 45 cm H_2O) requiring large minute volumes (> 10 $L \cdot min^{-1}$). Another imperfection of anesthesia ventilators is that they are pneumatically powered by compressed gases. The ventilator's rate of compressed gas consumption, which is approximately equal to the set minute volume (5–10 $L \cdot min^{-1}$ in a normal size adult), is not a concern when central compressed gas supplies are being used. But the ventilator can rapidly deplete oxygen supplies when compressed gas is being dispensed from the emergency backup cylinders attached to the anesthesia machine (e.g., a backup cylinder could provide over 10 h of oxygen to a breathing circuit at low flow, but would last only one-hour if also powering the ventilator). Lastly, anesthesia ventilators that do not sense patient effort are unable to provide synchronized or supportive modes of ventilation. This limitation is most significant during spontaneous ventilation, since CPAP and pressure support cannot be provided to compensate

for the additional work of breathing imposed by the breathing circuit and endotracheal tube, or to prevent the low lung volumes and atalectasis that result from general anesthesia. New anesthesia ventilators, introduced in the past 10 years, address many of these limitations as discussed later in the section on New Technologies.

Scavenger System

Waste anesthetic gases are vented from the operating room to prevent potentially adverse effects on health care workers. High volatile anesthetic concentrations in the operating room atmosphere can cause problems such as headaches, dysphoria, and impaired psychomotor functioning; chronic exposure to trace levels has been implicated as a causative factor for cancer, spontaneous abortions, neurologic disease, and genetic malformations, although many studies have not borne out these effects. The National Institute for Occupational Safety and Health (NIOSH) recommends that operating room levels of halogenated anesthetics be < 2 parts per million (ppm) and that nitrous oxide levels be < 25 ppm. Waste gases can be evacuated from the room actively via a central vacuum system, or passively via a hose to the outside; alternatively, the waste gas can pass through a canister containing activated charcoal, which absorbs halogenated anesthetics.

The scavenger system is the interface between the evacuation systems described in the preceding sentence and the exhaust valves on the breathing circuit and ventilator (i.e., APL valve and ventilator exhaust valve). It functions as a reservoir that holds waste gas until it can vent to the evacuation system. This is necessary because gas exits the exhaust valves at a non-constant rate that may, at times, exceed the flow rate of the evacuation system. The scavenger system also ensures that the downstream pressure on the exhaust valves does not become too high or too negative. Excessive pressure at the exhaust valve outlet could cause sustained high airway pressure leading to barotrauma and cardiovascular collapse; whereas, excessive vacuum at the exhaust valve outlet could cause sustained negative airway pressure leading to apnea and pulmonary edema.

There are two categories of scavenger systems, open and closed. Open scavenger systems can only be used with a vacuum evacuation system. In an open scavenger system, waste gas enters the bottom of a rigid reservoir that is open to the atmosphere at the top, and gas is constantly evacuated from the bottom of the reservoir into the vacuum. Room air is entrained into the reservoir whenever the vacuum flow rate exceeds the waste gas flow rate, and gas spills out to the room through the openings in the reservoir whenever the waste gas flow rate exceeds the vacuum flow rate. The arrangement of the components prevents spillage of waste gas out of the reservoir openings unless the *average* vacuum flow rate is less than the *average* flow out of the exhaust valves.

Closed scavenger systems consist of a compliant reservoir bag with an inflow of waste gas from the exhaust valves of the breathing system and an outflow to the active or passive evacuation system. Two or more valves regulate the internal pressure of the closed scavenger system. A negative pressure release valve opens to allow entry of room air whenever the pressure within the system becomes too negative, < -1.8 cm H_2O (-0.18 kPa) (i.e., in situations where the evacuation flow exceeds the exhaust flow and the reservoir bag is collapsed). A positive pressure release valve opens to allow venting of waste gas to the room whenever the pressure within the scavenger system becomes too high, > 5 cm H_2O (0.5 kPa) (i.e., in situations where the reservoir bag is full and the exhaust flow exceeds the evacuation flow). Thus, the pressure within the scavenger system is maintained between -1.8 and 5.0 cm H_2O.

Integrated Monitors

All anesthesia delivery systems have integrated electronic safety monitors intended to avert patient injuries. Included are (*1*) an oxygen analyzer, (*2*) an airway pressure monitor, and (*3*) a spirometer.

The oxygen analyzer measures oxygen concentration in the inspiratory limb of the breathing circuit to guard against the administration of dangerously low inhaled oxygen concentrations. Most analyzers use a polarographic or galvanic (fuel cell) probe that senses the rate of an oxygen-dependent electrochemical reaction. These analyzers are inexpensive and reliable, but are slow to equilibrate to changes in oxygen concentration (response times on the order of 30 s). They also require daily calibration. Standards stipulate that the oxygen analyzer be equipped with an alarm, and be powered-on whenever the anesthesia delivery system is in use.

The airway pressure monitor measures pressure within the breathing circuit, and warns of excessively high or negative pressures. It also guards against apnea during mechanical ventilation. Most anesthesia delivery systems have two pressure gauges: an analog Bourdon tube pressure gauge that displays instantaneous pressure on a mechanical dial, and an electronic strain-gauge monitor that displays a pressure waveform. Most electronic pressure monitors embody an alarm system with variable-threshold negative pressure, positive pressure, and sustained pressure alarms that can be adjusted by the user. An apnea alarm feature, which is enabled whenever the ventilator is powered-on, ensures that positive pressure is sensed within the breathing circuit at regular intervals. On some anesthesia delivery systems pressure is sensed within the circle system absorber canister; on other systems it is sensed on the patient side of the one-way valves; the latter gives a more accurate reflection of airway pressure.

The spirometer measures gas flow in the expiratory limb of the breathing circuit and guards against apnea and dangerously low or high respiratory volumes. A number of different techniques are commonly used to measure flow. These include spinning vanes, rotating sealed spirometers, ultrasonic, and variable orifice differential pressure. Respiratory rate, tidal volume, and minute volume are derived from the sensor signals and displayed to the user. Some machines also display a waveform of exhaled flow versus time. Most spirometers have an alarm system with variable-threshold alarms for low and high tidal volume, as well as an apnea alarm that is triggered if no flow is detected during a preset interval.

In addition to these standard monitors, some anesthesia workstations have integrated gas analyzers that measure inhaled and exhaled concentrations of oxygen, carbon dioxide, nitrous oxide, and volatile anesthetic agents. Although stand-alone gas analyzers are available, they are likely to be integrated into the anesthesia workstation because they monitor gas concentrations and respiratory parameters that are controlled by the anesthesia delivery system.

Other patient monitors, such as electrocardiography, pulse oximetry, invasive and noninvasive blood pressure, and thermometry may also be integrated into the anesthesia workstation; but often stand-alone monitors are placed on the shelves of the anesthesia delivery system. In either case, standard patient monitors must be used during the conduct of any anesthetic to evaluate the adequacy of the patient's oxygenation, ventilation, circulation, and body temperature. Monitoring standards, which have contributed to the dramatic increase in anesthesia safety, were initially published by the American Society of Anesthesiologists in 1986 and have been continually evaluated and updated (8).

New Technologies

The anesthesia delivery system as described thus far has evolved incrementally from a pneumatic device designed in 1917, by Henry Boyle for administration of anesthesia using oxygen, nitrous oxide and ether. The evolution of Boyle's machine has occurred in stages. In the 1950s and 1960s the failsafe devices and fluidic controlled ventilators were added. In the 1970s and early 1980s the focus was on improving safety with features, such as gas proportioning systems, safety alarms, electronically controlled ventilators, and standardization of the user interface to decrease errors. In the late 1980s and 1990s, monitors and electronic recordkeeping were integrated to create anesthesia workstations. Since 2000 the focus has been on improving ventilator performance, incorporating automated machine self-checks, and transitioning to electronically controlled and monitored flow meters and vaporizers. Some of the new technologies that have been introduced in the last few years are discussed, below.

BREATHING CIRCUIT

As discussed above, the tidal volume set on an anesthesia ventilator is not accurately delivered to the patient because of two breathing circuit effects. First, a portion of the volume delivered from the ventilator is compressed within the breathing circuit and does not reach the patient. Second, fresh gas flowing into the breathing circuit augments the delivered tidal volume. A number of techniques are used to minimize these effects in new anesthesia delivery systems.

Two techniques have been used to minimize the effect of gas compression. First, smaller, less compliant breathing circuits are being used. This has been achieved by minimizing the use of compliant hoses between the ventilator and breathing circuit and by decreasing the size of the absorber canister. A tradeoff is that the absorbent must be changed more frequently with a smaller canister, hence new breathing circuits are designed so that the carbon dioxide absorbent can be exchanged during use. Second, many new machines automatically measure breathing circuit compliance during an automated preuse checkout procedure and then compensate for breathing circuit compliance during positive pressure ventilation; the ventilator continually senses airway pressure and delivers additional volume to make up for that lost to compression.

A number of techniques have also been used to eliminate augmentation of tidal volume by fresh gas flowing into the circuit. In one approach, the ventilator automatically adjusts its delivered volume to compensate for the influx of fresh gas into the breathing circuit. The ventilator either adjusts to maintain a set exhaled tidal volume as measured by a spirometer in the expiratory limb of the breathing circuit, or it responds to maintain a set inhaled tidal volume sensed in the inspiratory limb, or it modifies its delivered volume based on the total fresh gas flow as measured by electronic flowmeters in the anesthesia machine. None of the above methods requires redesign of the breathing circuit, except for the addition of flow sensors that communicate with the ventilator.

In a radically different approach, called fresh gas decoupling, the breathing circuit is redesigned so that fresh gas flow is channeled away from ventilator-delivered gas during inspiration, which removes the augmenting effect of fresh gas flow on tidal volume. An example of such a breathing circuit is illustrated in Fig. 7. In this circuit, during inhalation, gas dispensed from a piston driven ventilator travels directly to the patient's lungs; retrograde flow is blocked by a passive fresh gas decoupling valve, and expiratory flow is blocked by the ventilator-controlled expiratory valve, which is actively closed during the inspiratory phase. Fresh gas does not contribute to the delivered tidal volume; instead it flows retrograde into a nonpressurized portion of the breathing circuit. During exhalation, the ventilator-controlled expiratory valve opens, and the ventilator piston withdraws to actively fill with a mixture of fresh gas and gas from the reservoir bag. This design causes a number of other functional changes. First, the breathing circuit compliance is lower during positive pressure ventilation, since only part of the breathing circuit is pressurized during inspiration (the volume between the fresh gas decoupling valve and the ventilator-controlled expiratory valve). Second, the reservoir bag remains in the circuit during mechanical ventilation. As a result, it fills and empties with gas throughout the ventilator cycle, which is an obvious contrast to the absence of bag movement during mechanical ventilation with a conventional circle breathing circuit.

ANESTHESIA MACHINE

Many new anesthesia machines have electronic gas flow sensors instead of tapered glass tubes with internal floats. Advantages include (*1*) improved reliability and reduced maintenance; (*2*) improved precision and accuracy at low-flows; and (*3*) ability to automatically record and use gas flows (for instance to adjust the ventilator). The electronic sensors operate on the principle of heat transfer, measuring the energy required to maintain the temperature of a heated element in the gas flow pathway. Each sensor is calibrated for a particular gas, since every gas has a different specific heat index. Gas flows are shown on dedicated light-emitting diode (LED) displays or on the

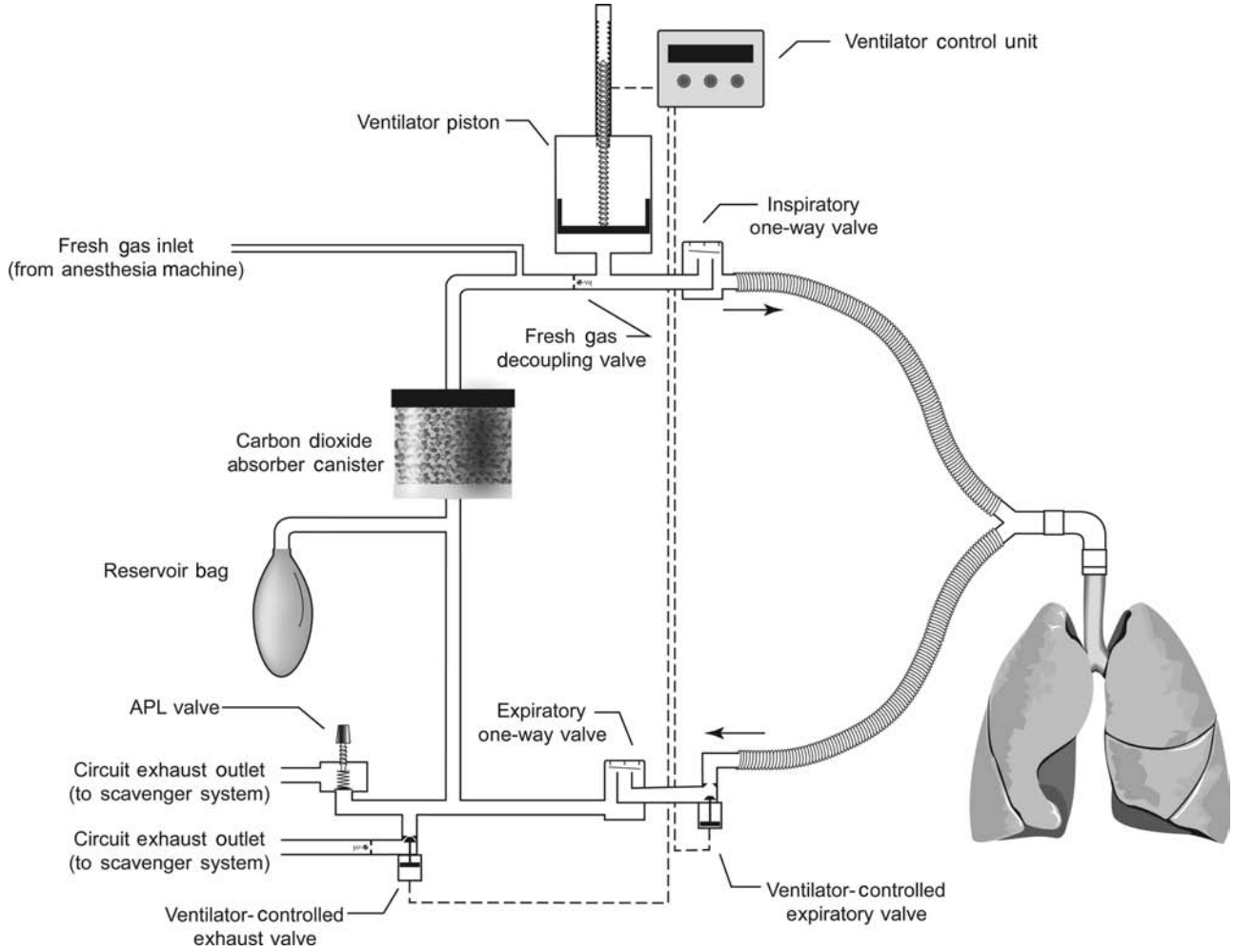


Figure 7. Example of a breathing circuit with fresh gas decoupling. This breathing circuit is used in the Draeger Fabius anesthesia machine. It contains three passive one-way valves and two active valves that are controlled by the ventilator during mechanical ventilation.

main anesthesia machine flat panel display. Most anesthesia machines still regulate the flow of each gas using mechanical needle valves, but in some these have been replaced with electronically control valves. Electronically controlled valves provide a mechanism for computerized gas proportioning systems that limit the ratios of multiple gases. Some machines with electronic flow control valves allow the user to select the balance gas (i.e., air or nitrous oxide) and set a desired oxygen concentration and total flow, leaving the calculation of individual gas flow rates to the machine.

Most new anesthesia machines continue to use mechanical vaporizers as described above, but a few incorporate electronic vaporizers. These operate on one of two principles: either computer-controlled variable bypass, or computer-controlled liquid injection. Computerized variable bypass vaporizers control an electronic valve that regulates the flow of gas exiting from the liquid anesthetic containing chamber to join the bypass stream. The valve is adjusted to reach a target flow that is based upon the: (*1*) dial setting, (*2*) temperature in the vaporizing chamber, (*3*) total pressure in the vaporizing chamber, (*4*) bypass flow, and (*5*) liquid anesthetic identity. Computerized injectors continuously add a measured amount liquid anesthetic directly into the mixed gas coming from the flowmeters based upon the: (*1*) dial setting, (*2*) mixed gas flow, and (*3*) liquid anesthetic identity. Electronic vaporizers offer a number of advantages. First, they provide a mechanism for vaporizer settings to be automatically recorded and controlled. Second, a number of different anesthetics can be dispensed (one at a time) using a single control unit, provided that the computer knows the identity of the anesthetic liquid.

VENTILATOR

Anesthesia ventilator technology has improved dramatically over the past 10 years and each new machine brings further advancements. As discussed above, most new ventilators compensate for the effects of circuit compliance and fresh gas flow, so that the set tidal volume is accurately delivered. Older style ventilators notoriously delivered low tidal volumes to patients requiring high airway pressures, but new ventilators overcome this problem with better flow generators, compliance compensation and feedback

control. Many new anesthesia ventilators offer multiple modes of ventilation (in addition to the traditional volume control), such as pressure control, pressure support, and synchronized intermittent mandatory ventilation. These modes assess patient effort using electronic flow and pressure sensors that are included in many new breathing circuits. Lastly, some anesthesia ventilators use an electronically controlled piston instead of the traditional pneumatically compressed bellows. Piston ventilators, which are electrically powered, dramatically decrease compressed gas consumption of the anesthesia delivery system. However, they actively draw gas out of the breathing circuit during the expiratory cycle (as opposed to bellows, which fill passively) so they cannot be used with a traditional circle system (see Fig. 7 for an example of a piston ventilator used with a fresh gas decoupled breathing circuit).

AUTOMATED CHECKOUT

Many new anesthesia delivery systems feature semiautomated preuse checkout procedures. These ensure that the machine is functioning properly prior to use by (*1*) testing electronic and computer performance, (*2*) calibrating flow sensors and oxygen monitors, (*3*) measuring breathing circuit compliance and leakage, and (*4*) testing the ventilator.

Future Challenges

The current trend is to design machines that provide advanced capabilities through the use of computerized electronic monitoring and controls. This provides the infrastructure for features such as closed-loop feedback, smart alarms, and information management that will be increasingly incorporated in the future. We can anticipate closed-loop controllers that automatically maintain a user-set exhaled anesthetic concentration (an indicator of anesthetic depth), or exhaled carbon dioxide concentration (an indicator of adequacy of ventilation). We can look forward to smart alarms that pinpoint the location of leaks or obstructions in the breathing circuit, alert the user and switch to a different anesthetic when a vaporizer becomes empty, or notify the user and switch to a backup cylinder if a pipeline failure or contamination event is detected. We can foresee information management systems that automatically incorporate anesthesia machine settings into a nationwide repository of anesthesia records that facilitate outcomes-guided medical practice, critical event investigations, and nationwide access to patient medical records. Anesthesia machine technology continues to evolve.

BIBLIOGRAPHY

Cited References

1. ASTM F1850. Standard Specification for Particular Requirements for Anesthesia Workstations and Their Components. ASTM International; 2000.
2. ISO 5358. Anaesthetic machines for use with humans. International Organization for Standardization; 1992.
3. ISO 8835-2. Inhalational anaesthesia systems—Part 2: Anaesthetic breathing systems for adults. International Organization for Standardization; 1999.
4. ISO 8835-3. Inhalational anaesthesia systems—Part 3: Anaesthetic gas scavenging systems—Transfer and receiving systems. International Organization for Standardization; 1997.
5. ISO 8835-4. Inhalational anaesthesia systems—Part 4: Anaesthetic vapour delivery devices. International Organization for Standardization; 2004.
6. ISO 8835-5. Inhalational anaesthesia systems—Part 5: Anaesthetic ventilators. International Organization for Standardization; 2004.
7. Anesthesia Apparatus Checkout Recommendations. United States Food and Drug Administration. Available at http://www.fda.gov//cdrh/humfac//anesckot.html. 1993.
8. Standards for Basic Anesthetic Monitoring, American Society of Anesthesiologists. Available at http://www.asahq.org/publicationsAndServices/standards/02.pdf. Accessed 2004.
9. CGA V-5. Diameter Index Safety System (Noninterchangeable Low Pressure Connections for Medical Gas Applications). Compressed Gas Association; 2000.
10. CGA V-1. Compressed Gas Association Standard for Compressed Gas Cylinder Valve Outlet and Inlet Connections. Compressed Gas Association; 2003.

Reading List

Dorsch J, Dorsch S. Understanding Anesthesia Equipment. 4th ed. Williams & Wilkins; 1999.
Brockwell RC, Andrews JJ. Inhaled Anesthetic Delivery Systems. In: Miller RD. et al. editors. Miller's Anesthesia. 6th ed. Philadelphia: Elsevier Churchill Livingstone; 2005.
Ehrenwerth J, Eisenkraft JB, editors. Anesthesia Equipment: Principles and Applications. St. Louis: Mosby; 1993
Lampotang S, Lizdas D, Liem EB, Dobbins W. The Virtual Anesthesia Machine. http://vam.anest.ufl.edu/.

See also CONTINUOUS POSITIVE AIRWAY PRESSURE; EQUIPMENT ACQUISITION; EQUIPMENT MAINTENANCE, BIOMEDICAL; GAS AND VACUUM SYSTEMS, CENTRALLY PIPED MEDICAL; VENTILATORY MONITORING.

ANESTHESIA MONITORING. See MONITORING IN ANESTHESIA.

ANESTHESIA, COMPUTERS IN

LE YI WANG
HONG WANG
Wayne State University
Detroit, Michigan

INTRODUCTION

Computer applications in anesthesia patient care have evolved with advancement of computer technology, information processing capability, and anesthesia devices and procedures.

Anesthesia is an integral part of most surgical operations. The objectives of anesthesia are to achieve hypnosis (consciousness control), analgesia (pain control), and immobility (body movement control) simultaneously throughout surgical operations, while maintaining the vital functions of the body. Vital functions, such as respiration and circulation of blood, are assessed by signs such as blood pressures, heart rate, end-tidal carbon dioxide (CO_2), oxygen saturation by pulse oximetry (SpO_2), and so on. These objectives are

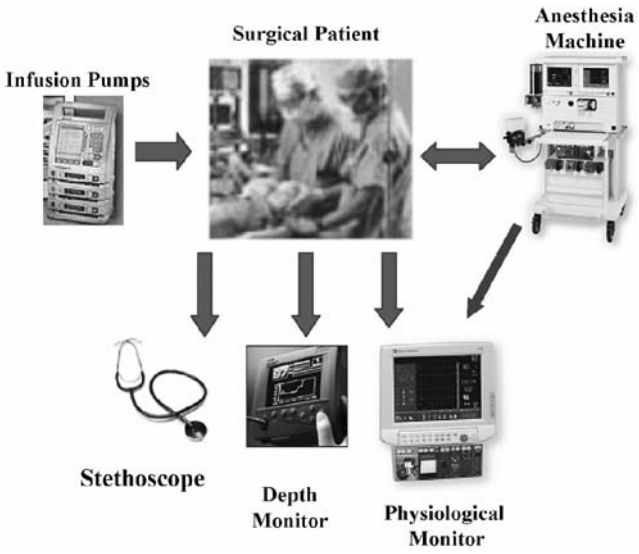


Figure 1. Some anesthesia equipment in an operating room.

carefully balanced and maintained by a dedicated anesthesia provider using a combination of sedative agents, hypnotic drugs, narcotic drugs, and, in many surgeries, muscle relaxants. Anesthesia decisions and management are complicated tasks, in which anesthetic requirements and agent dosages depend critically on the surgical procedures, the patient's medical conditions, drug interactions, and coordinated levels of anesthesia depth and physiological variables. Anesthesia decisions impact significantly on surgery and patient outcomes, drug consumptions, hospital stays, and therefore quality of patient care and healthcare cost (Fig. 1).

Computer technologies have played essential roles in assisting and improving patient care in anesthesia. Development of computer technology started from its early stages of bulky computing machines, progressed to minicomputers and microcomputers, exploded with its storage capability and computational speed, and evolved into multiprocessor systems, distributed systems, computer networks, and multimedia systems. Computer applications in anesthesia have taken advantage of this technology advancement. Early computer applications in medicine date back to the late 1950s when some hospitals began to develop computer data processing systems to assist administration, such as storage, management, and analysis of patient and procedural data and records. The main goals were to reduce manpower in managing ever-growing patient data, patient and room scheduling, anesthesia supply tracking, and billing. During the past four decades computer utility in anesthesia has significantly progressed to include computer-controlled fluid administration and drug dispersing, advanced anesthesia monitoring, anesthesia information systems, computer-assisted anesthesia control, computer-assisted diagnosis and decisions, and telemedicine in anesthesia.

COMPUTER UTILITY IN ADVANCED ANESTHESIA MONITORING

The quality of anesthesia patient care has been greatly influenced by monitoring technology development. A

(a) Paramagnetic Oxygen Analyzer in 1940 (b) Carbon Dioxide Monitor in 1960

Figure 2. Anesthesia monitoring devices without computer technologies. (Courtsey of Sheffield Museum of Anesthesia used with permission.)

patient's state during a surgery is assessed using vital signs. In earlier days of anesthesiology, vital signs were limited to manual measurements of blood pressures, stethoscope auscultation of heart–lung sounds, and heart rates. These values were measured intermittently and as needed during surgery. Thanks to advancement of materials, sensing methods, and signal processing techniques, many vital signs can now be directly, accurately, and continuously measured. For example, since the invention of pulse oximetry in the early 1980s, this noninvasive method of continuously monitoring the arterial oxygen saturation level in a patient's blood (SpO_2) has become a standard method in the clinical environment, resulting in a significant improvement of patient safety. Before this invention, blood must be drawn from patients and analyzed using laboratory equipment.

Integrating these vital signs into a comprehensive anesthesia monitoring system has been achieved by computer data interfacing, multisignal processing, and computer graphics. Advanced anesthesia monitors are capable of acquiring multiple signals from many vital sign sensors and anesthesia machine itself, displaying current readings and historic trends, and providing audio and visual warning signals. At present, heart rate, electrocardiogram (ECG), arterial blood pressures, temperature, ventilation parameters (inspired–expired gas concentration, peak airway pressure, plateau airway pressure, inspired and expired volumes, etc.), end-tidal CO_2 concentrations, blood oxygen saturation (SpO_2), and so on, are routinely and reliably monitored (Figs. 2 and 3).

However, there are still many other variables reflecting a patient's state that must be inferred by the physician, such as anesthesia depth and pain intensity. Pursuit of new

Figure 3. An anesthesia monitor from GE Healthcare in 2005. (Courtsey of GE Healthcare.)

physiological monitoring devices for direct and reliable measurements of some of these variables is of great value and imposes great challenges at the same time (1). Anesthesia depth has become a main focus of research in the anesthesia field. At present, most methods rely in part or in whole on processing of the electroencephalogram (EEG) and frontalis electromyogram (FEMG) signals. Proposed methods include the median frequency, spectral edge frequency, visual evoked potential, auditory evoked potential, entropy, and bispectral index (2,3). Some of these technologies have been commercialized, leading to several anesthesia depth monitors for use in general anesthesia and sedation. Rather than using indirect implications from blood pressures, heart rate, and involuntary muscle movements to derive consciousness levels, these monitors purport to give a direct index of a patient's anesthesia depth. Consequently, combined effects of anesthesia drugs on the patient anesthesia depth can potentially be understood clearly and unambiguously. Currently (the year 2005), the BIS Monitor by Aspect Medical Systems, Inc. (www.aspectmedical.com), Entropy Monitor by GE Healthcare (www.gehealthcare.com), and Patient State Analyzer (PSA) by Hospira, Inc. (www.hospira.com) are three FDA (U.S. Food and Drug Administration) approved commercial monitors for anesthesia depth.

Availability of commercialized anesthesia depth monitors has prompted a burst of research activity on computerized depth control. Improvement of their reliability remains a research frontier. Artifacts have fundamental impact on reliability of EEG signals. In particular, muscle movements, eye blinks, and other neural stimulation effects corrupt EEG signals, challenging all the methods that rely on EEG to derive anesthesia depth. As a result, reliability of these devices in intensive care units (ICU) and emergency medicine remains to be improved.

Another area of research is pain-intensity measurement and monitoring. Despite a long history of research and development, pain intensity is still evaluated by subjective assessment and patient self-scoring. The main thrust is to establish the relation between subjective pain scores, such as the visual analog scale (VAS) system, and objective measures of vital signs. Computer-generated objective and continuous monitoring of pain will be a significant advance in anesthesia pain control. This remains an open and active area of research and development (R&D). As an intermediate step, patient-controlled analgesia (PCA) devices have been developed (see, e.g., LifeCare PCA systems from Hospira, Inc., which is a 2003 spin-off of Abbott Laboratories) that allow a patient to assess his/her pain intensity and control analgesia as needed.

Currently, anesthesia monitors are limited to data recording and patient state display. Also, their basic functions do not offer substantial interaction with human and environment. Future monitors must enhance fundamentally human-factors design: Intelligent human–machine interface and integrated human–machine–environment systems (4). Ideally, a monitor will intelligently organize observation data into useful information, adapt its functions according to surgical and anesthesia events, select the most relevant information to display, modify its display layouts to reduce distraction and amplify essential information, tune safety boundaries for its warning systems on the basis of the individual medical conditions of the patient, analyze data to help diagnosis and treatment, and allow user-friendly interactive navigation of the monitor system. Such a monitor will eventually become an extension of a physician's senses and an assistant of decision-making processes.

COMPUTER INFORMATION TECHNOLOGY IN ANESTHESIA

Anesthesia Information Systems

Patient information processing systems have undergone a long history of evolution. Starting in the 1960s, some computer programming software and languages were introduced to construct patient information systems. One example is MUMPS (Massachusetts General Hospital Utility Multi-Programming System), which was developed in Massachusetts General Hospital and used by other hospitals, as well as the U.S. Department of Defense and the U.S. Veteran's Administration. During the same period, Duke University's GEMISCH, a multi-user database programming language, was created to streamline data sharing and retrieval capabilities.

Currently, a typical anesthesia information system (AIS) consists of a central computer station or a server that is interconnected via wired or wireless data communication networks to many subsystems. Subsystems include anesthesia monitors and record-keeping systems in operating rooms, preoperative areas, postanesthesia care units (PACU), ICUs; data entry and record systems of hospital testing labs; office computers of anesthesiologists. The system also communicates with hospital mainframe information systems to further exchange information with in- and out-patient care services, patient database, and billing systems.

Information from an operating room is first collected by medical devices and anesthesia monitors and locally sorted and recorded in the record-keeping system. Selected data are then transmitted to the mainframe server through the data network. Anesthesia events, procedures, physician observations and diagnosis, patient care plans, testing results, drug and fluid data can also be entered into the record-keeping system, and broadcast to the main server and/or other related subsystems.

The main server and observation station provide a center in which patient status in many operating rooms, preoperative area and PACUs can be simultaneously monitored in real time. More importantly, the central anesthesia information system keeps accurately patient data and makes them promptly accessible to many important functions, including patient care assessment, quality assurance, room scheduling, physician assignment, clinical studies, medical billing, regulation compliance, equipment and personnel utility, drug and blood inventory, postoperative in-patient and out-patient service, to name just a few examples (Fig. 4).

One example of AIS is the automation software system CareSuite of PICIS, Inc. (www.picis.com). The system delivers comprehensive perioperative automation. It provides surgical

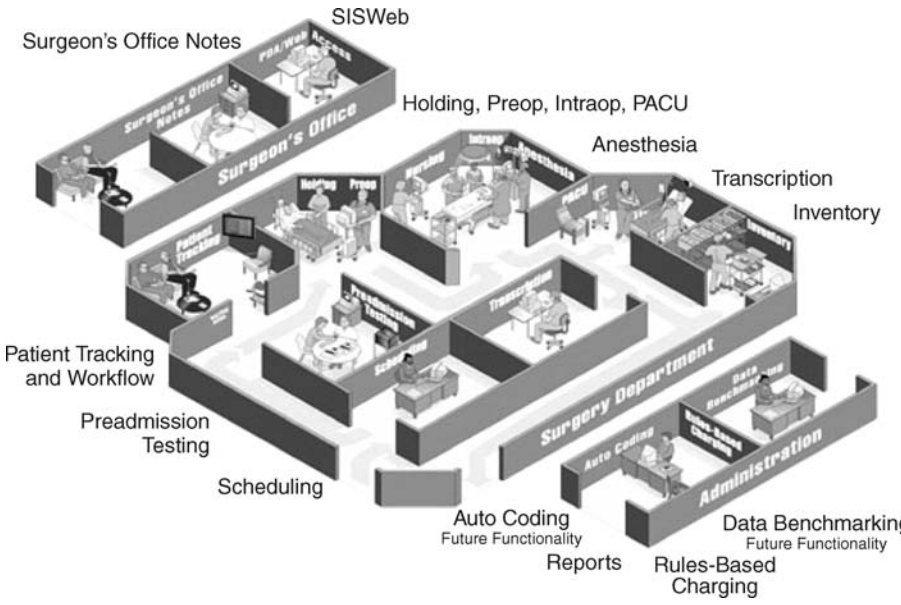


Figure 4. An illustration of a surgical/anesthesia information management system from surgical information systems, Inc. (www. orsoftware.com).

and anesthesia supply management, intraoperative nursing notes, surgical infection control monitoring, adverse event tracking and intervention, patient tracking, resource tracking, outlier alerts, anesthesia record, anesthesia times, case compliance, and so on. Similarly, the surgical and anesthesia management software by Surgical Information Systems (SIS), Inc. (www.orsoftware.com) streamlines patient care and facilitates analysis and performance improvement.

Anesthesia information systems are part of an emerging discipline called medical informatics, which studies clinical information and clinical decision support systems. Although technology maturity of computer hardware and software has made medical information systems highly feasible, creating seamless exchange of information among disparate systems remain a difficult task. This presents new opportunities and challenges for broader application of medical informatics in anesthesia practice.

Computer Simulation: Human Patient Simulators

Human patient simulators (HPS) are computerized mannequins whose integrated mechanical systems and computer hardware, and sophisticated computer software mimic authentically many related physiological, pathological, and pharmacological aspects of the human patient during a surgery or a medical procedure (Fig. 5). The mannequins are designed to simulate an adult or pediatric patient of either gender under a medical stress condition. They are accommodated in a clinical setting, such as an operating room, a trauma site, an emergence suite, or an ICU. Infusion pumps use clean liquid (usually water) through bar-coded syringes, whose barcodes are read by the HPS code recognition system to identify the drugs, to administer the simulated drugs, infusion liquids, or transfused blood during an anesthesia administration. The HPS allows invasive procedures such as intubation.

The HPS responds comprehensively to administered drugs, surgical events, patient conditions, and medical crisis; and displays on standard anesthesia monitors most related physiological vital signs, including blood pressures, heart rate, EKG, and oxygen saturations. They also generate normal and adventurous heart and lung sounds for auscultation. All these characteristics are internally generated by the computer software that utilizes mathematics models of typical human patients to simulate the human responses. The patient's physical parameters (age, weight, smoker, etc.), preexisting medical conditions (high blood pressure, asthma, diabetic, etc.), surgical procedures, clinical events, and critical conditions are easily programmed by a computer Scenario Editor with a user-friendly graphical interface. The Scenario Editor also allows interactive reprogramming of scenarios during an operation. This on-the-fly function of scenario generation is especially useful for training. It gives the instructor great flexibility to create new scenarios according to the trainee's reactions to previous scenarios.

The HPS is a great educational tool that has been used extensively in training medical students, nurse anesthetists, anesthesia residents, emergency, and battlefield

Figure 5. Human patient simulator complex at Wayne State University.

Figure 6. Anesthesia resident training on an HPS manufactured by METI, Inc. (Used with permission.)

medics. Its preliminary development can be traced back to the 1950s, with limited computer hardware or software. Its more comprehensive improvement occurred in pace with computer technology in the late 1960s when highly computerized models were incorporated into high fidelity HPS systems with interfaces to external computers.

Due to rareness of anesthesia crisis, student and resident training on frequent and repeated critical medical conditions and scenarios is not possible in operating rooms. The HPS permits the trainee to practice clinical skills and manage complex and critical clinical conditions by generating and repeating difficult medical scenarios. The instructor can design individualized programs to evaluate and improve trainees' crisis management skills. For invasive skills, such as intubation, practice on simulators is not harmful to human patients. Catastrophic or basic events are presented with many variations so that trainees can recognize their symptoms, diagnose their occurrences, treat them according to established guidelines, and avert disasters. For those students who have difficulties to transform classroom knowledge to clinical hands-on skills, the HPS training is a comfortable bridge for them to rehearse in simulated clinical environments (5) (Fig. 6).

There are several models of HPSs on market. For example, the MedSim-Eagle Patient Simulator (Fig. 7) is a realistic, hands-on simulator of the anesthetized or critically ill patient, developed at Stanford University and manufactured by Eagle Simulation, Inc. METI (Medical Education Technologies, Inc.) (www.meti.com) manufactures adult HPS (Stan), pediatric HPS (PediaSim), emergency care simulator (ECS), pediatric emergency simulator (PediaSim-ECS), and related simulation suites such as airway tools (AirSim), surgical training tools (SurgicalSim). Laerdal Medical AS (www.laerdal.com) has developed a comprehensive portable HPS that can be operated without the usual operating room settings for HPS operations.

Human simulations, however, are not a total reality. Regardless how comprehensive the HPS has become, real environments are far more complex. There are many complications that cannot be easily simulated. Consequences of overly aggressive handling of certain medical catastrophic events may not be fully represented. Issues like these have prompted further efforts in improving HPS technologies and enhancing their utilities in anesthesia education, training, and research.

Figure 7. A human patient simulator SimMan, by Laerdal Medical Corporation. (Used with permission.)

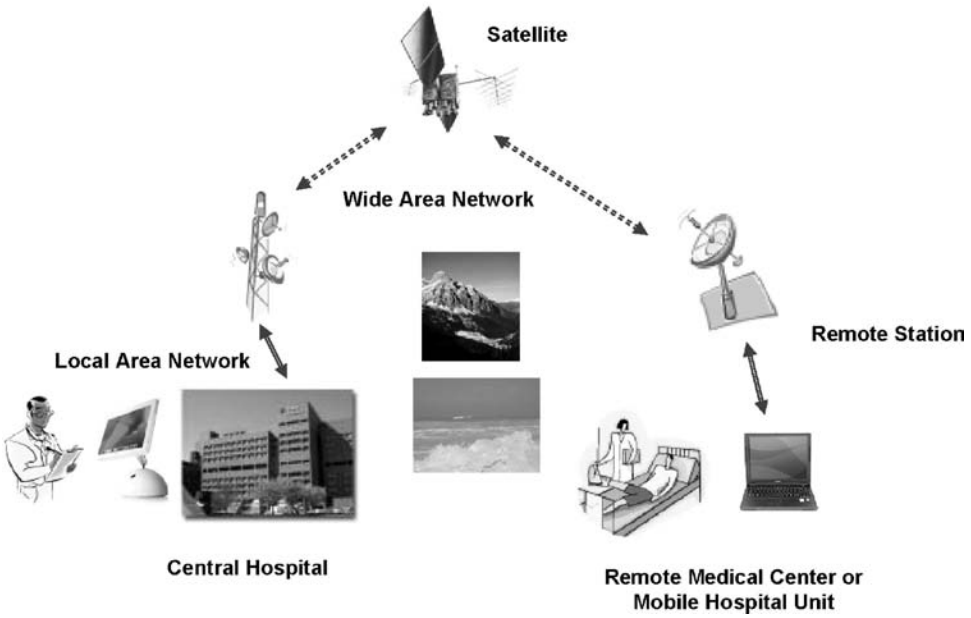


Figure 8. Telemedicine connects remote medical centers for patient care.

Large Area Computer Networks: Telemedicine in Anesthesia

Telemedicine can be used to deliver healthcare over geographically separated locations. High speed telecommunication systems allow interactive video-mediated clinical consultation, and the possibility in the future of remote anesthesia administration. Wide availability of high speed computer and wireless network systems have made telemedicine a viable area for computer applications. Telemedicine could enable the delivery of specialized anesthesia care to remote locations that may not be accessible to high quality anesthesia services and knowledge, may reduce significantly travel costs, and expand the supervision capability of highly trained anesthesiologists.

In a typical telemedicine anesthesia consultation, an anesthesiologist in the consultation center communicates by a high speed network with the patient and the local anesthesia care provider, such as a nurse, at the remote location (Fig. 8). Data, audio and video connections enable the parties to transfer data, conduct conversations on medical history and other consultation routines, share graphs, discuss diagnosis, and examine the patient by cameras. The anesthesiologist can evaluate the airway management, ventilation systems, anesthesia monitor, and cardiovascular systems. Heart and lung sound auscultation can be remotely performed. Airway can be visually examined. The anesthesiologist can then provide consultation and instructions to the remote anesthesia provider on anesthesia management.

Although telemedicine is a technology-ready field of computer applications and has been used in many medicine specialties, at present its usage for systematic anesthesia consultation remains at its infancy. One study reports a case of telemedicine anesthesia between the Amazonian rainforests of Ecuador and Virginia Commonwealth University, via a commercially developed telemedicine system (6). In another pilot study, the University Health Network in Toronto utilized Northern Ontario Remote Telecommunication Health (NORTH) Network to provide telemedicine clinical consultations to residents of central and northern Ontario in Canada (7).

COMPUTER-AIDED ANESTHESIA CONTROL

The heart of most medical decisions is a clear understanding of the outcome from drug administration or from specific procedures performed on the patient. To achieve a satisfactory decision, one needs to characterize outcomes (outputs), establish causal links between drugs and procedures (inputs) and the outcomes, define classes of decisions in consideration (classes of possible actions and controllers), and design actions (decisions and control). Anesthesia providers perform these cognitive tasks on the basis of their expertise, experience, knowledge of guidelines, and their own subjective judgments. It has long been perceived in the field of anesthesiology that computers may help in this decision and control process.

At a relatively low level of control and decision assistance, there has been routine use by anesthesia providers of computers to supply comprehensive and accurate information about anesthesia drugs, procedures, and guidelines in relation to individual patient care. Thanks to miniaturization and internets, there are now commonly and commercially available digital reference databases on anesthesia drugs, their detailed user manuals, and anesthesia procedures. With a palm-held device, all information becomes readily available to anesthesia providers in operating rooms, and other clinical settings. New data can be routinely downloaded to keep information up-to-date.

More challenging aspects of computer applications are those involving uncertainty, control, and intelligence that are the core of medical decision processes. These include individualized models of human patients, outcome prediction, computer-assisted control, diagnosis, and decision assistance. Such tools need to be further developed and commercialized for anesthesia use.

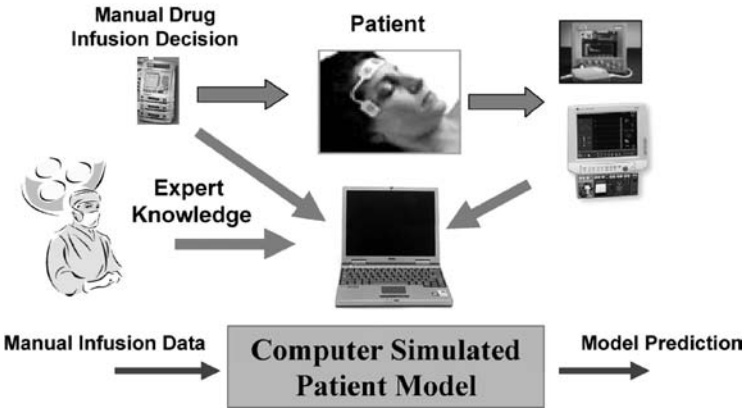


Figure 9. Utility of patient models to predict outcomes of drug infusion.

Patient Modeling and Outcome Prediction

Response of a patient's physiological and pathological state to drugs and procedures is the key information that an anesthesiologist uses in their management. The response, that is, the outcome, can be represented by either the values of the patient vital signs such as anesthesia depth and blood pressures, or consequence values such as length of ICU stay, hospital stay, complications. Usually, drug impact on patient outcomes is evaluated in clinical trials on a large and representative population and by subsequent statistical analysis. These population-based models (average responses of the selected population) link drug and procedural inputs to their effects on the patient state. These models can then be used to develop anesthesia management guidelines (Fig. 9).

For real-time anesthesia control in operating rooms, the patient model also must represent dynamic aspects of the patient response to drugs and procedures (8). This real-time dynamic outcome prediction requires a higher level of modeling accuracy, and is more challenging than off-line statistical analysis of drug impact. Real-time anesthesia control problems are broadly exemplified by anesthesia drug infusion, fluid resuscitation, pain management, sedation control, automated drug rates for diabetics, and so on.

There have been substantial modeling efforts to capture pharmacokinetic and pharmacodynamic aspects of drug impact as well as their control applications (9). These are mostly physiology-based and compartment-modeling approaches. By modeling each process of infusion pump dynamics, drug propagation, concentration of drugs on various target sites, effect of drug concentration on nerve systems, physiological response to nerve stimulations, and sensor dynamics, an overall patient response model can be established. Verification of such models has been performed by comparing model-predicted responses to measured drug concentration and physiological variables. These models have been used in evaluating drug impact, decision assistance and control designs.

Computer Automation: Anesthesia Control Systems

At present, an anesthesiologist decides on an initial drug control strategy by reviewing the patient's medical conditions, then adapts the strategy after observing the patient's actual response to the drug infusion. The strategy is further tuned under different surgical events, such as incision, operation, and closing. Difficulties in maintaining smooth and accurate anesthesia control can have dire consequences, from increased drug consumption, side effects, short- and long-term impairments, and even death. Real-time and computer-assisted information processing can play a pivotal role in extracting critical information, deriving accurate drug outcome predictions, and assisting anesthesia control.

Research efforts to develop computer-assisted anesthesia control systems have been ongoing since the early 1950s (10–14). The recent surge of interest in computer-assisted anesthesia diagnosis, prediction, and controls is partly driven by the advances in anesthesia monitoring technologies, such as depth measurements, computer-programmable infusion pumps, and multisignal real-time data acquisition capabilities. These signals provide fast and more accurate information on the patient state, making computer-aided control a viable possibility. Research findings from computer simulations, animal studies, and limited human trials, have demonstrated that many standard control techniques, such as proportional-integral-derivative (PID) controllers, nonlinear control techniques, fuzzy logic, model predictive control, can potentially provide better performance under routine anesthesia conditions in operating rooms (Fig. 10).

Target Concentration and Effect Control. Target concentration or drug effect control is an open-loop control strategy. It relies on computer models that relate drug infusion rates to drug concentrations on certain target sites or to drug effects on physiological or nerve systems. Since at present drug concentration or drug effects are not directly measured in real-time, feedback control is often not possible. Implementation of this control strategy can be briefly described as follows. For a prespecified time interval, the desired drug concentration profile is defined. This profile is usually determined *a priori* by expert knowledge, safety mandates, and smooth control requirements. A performance index is then devised that includes terms for control accuracy (to follow the desired profiles closely), drug consumption, constraints on physiological variables (safety constraints), and so on. Then, an optimal control is derived by optimizing the performance index under the given constraints and the dynamic

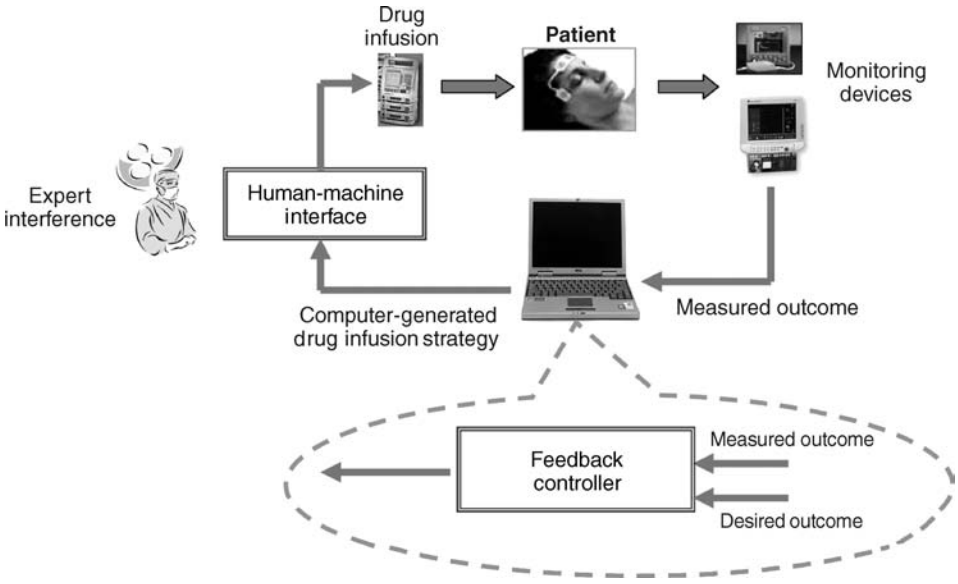


Figure 10. Computer-assisted drug infusion control (a) without expert interference: automated anesthesia feedback control. (b) With expert interference: anesthesia decision assistance systems.

models of the patient. One common method of designing optimal control strategies is dynamic programming, although many other optimal or suboptimal control design methodologies for nonlinear control systems are also available in the control field. Due to lack of feedback correction in target concentration control, optimality and accuracy of control actions may be compromised. However, even this open-loop control has seen many successful applications, such as glucose level control. Feedback control may become feasible in the future when new sensors become available to measure directly drug concentration.

Automatic Feedback Control. Computer-assisted anesthesia control has been frequently compared to autopilot systems in aviation. The autopilot system controls flying trajectories, altitude and position, airplane stability and smoothness, automatically with minimum human supervision. Success of such systems has resulted in their ubiquitous applications in most airplanes. It was speculated that an anesthesia provider's routine control tasks during a surgery may be taken over by a computer that adjusts automatically drug infusions to maintain desirable patient states. Potential and speculated advantages of such systems may include less reliance on experience, avoidance of fatigue-related medical mistakes, smoother control outcomes, reduced drug consumptions, and consequently faster recovery. So far, these aspects have been demonstrated only in a few selective cases of research subjects.

System Identification and Adaptive Control for Individualized Control Strategies. One possible remedy for compensating variations in surgical procedures and patient conditions in control design is to use real-time data to adjust patient models that are used in either target concentration control or feedback control. Successful implementation of this idea will generate individualized models that will capture the unique characteristics of the patient. This real-time patient model can then be used to tune the controllers that will deliver best performance for the patient. This control tuning is the core idea of adaptive control systems: Modifying control strategies on the basis of individually identified patient models. Adaptive control has been successfully applied to a vast array of industrial systems. The main methods of model reference adaptive control, gain scheduling, self-tuning regulators, and machine learning are potentially applicable in anesthesia control. This is especially appealing since variations and uncertainties in patient conditions and surgical events and procedures are far more complicated than industrial systems.

Most control algorithms that have been employed in anesthesia control are standard. The main difficulties in applying automated anesthesia control are not the main control methodologies, but rather an integrated system with high reliability and robustness, and well-designed human-machine interaction and navigation. Unlike an airplane in midair or industrial systems, anesthesia patients vary vastly in their responses to drugs and procedures. Control strategies devised for a patient population may not work well in individual patients. Real-time, on-site, and automatic calibration of control strategies are far more difficult than designing an initial control strategy for a patient population. Adaptation adds a layer of nonlinear feedback over the underlying control, leading to adaptive PID, tuned fuzzy, adaptive neural frameworks, and so on. Stability, accuracy, and robustness of such control structures are more difficult to establish. Furthermore, human interference must be integrated into anesthesia control systems to permit doctors to give guidelines and sometimes take control. Due to high standard in patient safety, at present automated anesthesia control remains largely in a phase of research, and in a very limited sense, toward technology transfer to medical devices. It will require a major commercialization effort and large clinical studies to transform research findings into product development of anesthesia controllers.

Moreover, medical complications occur routinely, which cannot be completely modeled or represented in control

strategies. Since such events usually are not automatically measured, it is strenuous to compensate their impact quickly. In addition, medical liability issues have raised the bar of applying automated systems. These concerns have curtailed a widespread realization of automated anesthesia control systems, despite a history of active research over four decades on anesthesia control systems.

COMPUTER INTELLIGENCE: DIAGNOSIS AND DECISION ASSISTANCE

In parallel to development of automatic anesthesia control systems, a broader application of computers in anesthesia management is computer-aided anesthesia diagnosis, decision assistance, and expert systems. Surveys of anesthesia providers have indicated that the field of anesthesiology favors system features that advise or guide rather than control (15). Direct interventions, closed-loop control, lockout systems, or any other coercive method draw more concerns. In this aspect, it seems that anesthesia expert decision support systems may be an important milestone to achieve before automated systems.

Anesthesia Diagnosis and Decision Assistance

Computer-aided diagnosis will extract useful information from patient data and vital-sign measurements, apply computerized logic and rigorous evaluations of the data, provide diagnosis on probable causes, and suggest guideline-driven remedy solutions. The outcome of the analysis and diagnosis can be presented to the anesthesia care provider with graphical displays, interactive user interfaces, and audio and visual warnings.

Decision assistance systems provide decision suggestions, rather direct and automatic decision implementations. Such systems provide a menu of possible actions for an event, or dosage suggestions for control purposes, and potential consequences of selected decisions. Diagnosis of possible causes can remind the anesthesiologist what might be overlooked in a crisis situation. The system can have interactive interfaces to allow the physician to discuss further actions and the corresponding outcomes with the computer. This idea of physician-assistant systems aims to provide concise, timely, and accurate references to the anesthesiologist for improved decisions. Since the physician remains as the ultimate decision maker, their management will be enhanced by the available information and diagnosis, but not taken over.

Suggested remedies of undesirable events are essentially recommendations from anesthesia management guidelines, brought out electronically to the anesthesiologist. Some computer simulators for anesthesia education are developed on the basis of this idea. For example, Anesthesia Simulator by Anesoft Corporation (www.anesoft.com) contains a software module of expert consultation that incorporates anesthesia emergency scenarios and suggests expert advices. Utility of expert systems in resident training has been widely accepted. However, decision support systems in the operating rooms are slow in development and acceptance. Generally speaking, a decision support system must interact with the complicated cognitive environment of the operating rooms. To make such systems a useful tool, they must be designed to accommodate the common practice in which the anesthesiologist thinks, sees, and reasons, rather than imposing a complicated new monitoring mode for the clinician to be retrained. This is again an issue of human-factors design.

Dosage recommendations for anesthesia drugs are internally derived from embedded modeling and control strategies. In principle, the control strategies discussed in the previous sections can support the decision assistance system. By including the physician in the decision loop, some issues associated with automated control systems can be alleviated. Reliability of such control strategies, user interfaces, and clinical evidence of cost-effectiveness of the decision support system will be the key steps toward successful clinical applications of such systems.

FUTURE UTILITY OF COMPUTER TECHNOLOGY IN ANESTHESIA

The discussions in the previous sections outline briefly critical roles that computers have played in improving anesthesia management. New development in computer-related technologies are of much larger potential.

Micro-Electro-Mechanical Systems (MEMS) is a technology that integrates electrical and mechanical elements on a common silicon material. This technology has been used in developing miniature sensors and actuators, such as micro infusion pumps and *in vivo* sensors. Integrated with computing and communication capabilities, these devices become smart sensors and smart actuators. The MEMS technology has reached its maturity. Further into the realms of fabrication technology at atom levels, emergence of nanotechnology holds even further potential of new generations of medical devices and technologies. There are many exciting possibilities for utility of these technologies in anesthesia: *In vitro* sensors based on nanodevices can potentially pinpoint drug concentrations at specific target sites, providing more accurate values for automated anesthesia drug control; Microactuators can directly deliver drugs to the target locations promptly and accurately, reducing drastically reliance on trial-and-error and sharpened experience in anesthesia drug infusion control; MEMS and nanosensors together with computer graphical tools will allow two-dimensional (2D) or three-dimensional (3D) visual displays of drug propagation, drug concentration, distributed blood pressures, heart and lung functions, brain functions, consequently assisting anesthesiologists in making better decisions about drug delivery for optimal patient care.

On another frontier of technology advancement, computer parallel computing (many computers working in symphony to solve complicated problems), computer imaging processing, data mining (extracting useful information from large amount of data), machine intelligence, wireless communication technologies, and human-factors science and design provide a vast opportunity and a promising horizon in advancing anesthesia management.

Advanced anesthesia control systems will manage routine drug infusion with their control actions tuned to

individual patients' conditions and surgical procedures, relieving anesthesiologists from stressful and strenuous routine tasks to concentrate on higher level decisions in patient care.

Patient physiological conditions can be more accurately and objectively measured by computer-processed sensor and imaging information.

Computer-added imaging processing will make it possible to consolidate information from CT-Scan (Computed Tomography), TEE (Transesophageal Echocardiography), MRI (Magnetic Resonance Imaging), and fMRI (Functional Magnetic Resonance Imaging) into regular anesthesia monitoring.

Anesthesia decisions will be assisted by computer database systems and diagnosis functions.

Anesthesia monitoring devices will become wireless, eliminating the typical spaghetti conditions of monitoring cables in operating rooms.

Anesthesia information systems will become highly connected and standard in anesthesia services, automating and streamlining the total patient care system: From patient admission to patient discharge, as well as follow-up services.

BIBLIOGRAPHY

Cited References

1. Penelope M, et al. Advanced patient monitoring displays: tools for continuous informing. Anesthes Analges 2005;101: 161–168.
2. Bonhomme V, et al. Auditory steady-state response and bispectral index for assessing level of consciousness during propofol sedation and hypnosis. Intravenous Anesthes 2000; 91:1398–1403.
3. Drummond JC. Monitoring depth of anesthesia. Anesthesiology 2000;93(3):876–882.
4. Blike GT. Human factors engineering: It's all about 'usability'. ASA Newslett Oct. 2004;68.
5. Peteani LA. Enhancing clinical practice and education with high-fidelity human patient simulators. Nurse Educ 2004; 29(1):25–30.
6. Stephen W, et al. Case report of remote anesthetic monitoring using telemedicine. Anesthes Analges 2004;98:386–388.
7. Wong DT, et al. Preadmission anesthesia consultation using telemedicine technology: A pilot study. Anesthesiology June 2004;100(6):1605–1607.
8. Wang LY, Yin G, Wang H. Identification of Wiener models with anesthesia applications. Int J Pure Appl Math Sci 2004; 35–61.
9. Shafer A, Doze VA, Shafer SL, White PF. Pharmacokinetics and pharmacodynamics of propofol infusions during general anesthesia. Anesthesiology 1988;69:348–356.
10. Eisenach JC. Reports of scientific meetings-workshop on safe feedback control of anesthetic drug delivery. Anesthesiology August 1999;91:600–601.
11. Linkens DA, Hacisalihzade SS. Computer control systems and pharmacological drug administration: A survey. J Med Eng Technol 1990;14(2):41–54.
12. Mortier EM, et al. Closed-loop controlled administration of Propofol using bispectral analysis. Anaesthesia 1998;53: 749–754.
13. Rao RR, et al. Automaded regulation of hemodynamic variables. IEEE Eng Med Biol Mag 2001;20:24–38.
14. Tackley RM, et al. Computer controlled infusion of propofol. Br J Anesthes 1989;62:46–53.
15. Beatty PT, et al. User attitudes to computer-based decision support In anesthesia and critical care: A preliminary survey. Internet J Anesthesiol 1999;3(1).

See also ANESTHESIA MACHINES; CARDIAC OUTPUT, THERMODILUTION MEASUREMENT OF; ELECTROCARDIOGRAPHY, COMPUTERS IN; ELECTROENCEPHALOGRAPHY; MEDICAL RECORDS, COMPUTERS IN; MONITORING IN ANESTHESIA.

ANGER CAMERA

MARK T. MADSEN
University of Iowa

INTRODUCTION

In nuclear medicine, radioactive tracers are used to provide diagnostic information for a wide range of medical indications. Gamma-ray emitting radionuclides are nearly ideal tracers, because they can be administered in small quantities and yet can still be externally detected. When the radionuclides are attached to diagnostically useful compounds (1), the internal distribution of these compounds provides crucial information about organ function and physiology that is not available from other imaging modalities. The Anger camera provides the means for generating images of the radiopharmaceuticals within the body. Example images of some common studies are shown in Figs. 1 and 2.

Initially, nonimaging detectors were used to monitor the presence or absence of the radiotracer. However, it was clear that mapping the internal distribution of the radiotracers would provide additional diagnostic information. In 1950, Benedict Cassen introduced the rectilinear scanner. The rectilinear scanner generated images of radionuclide distributions by moving a collimated sodium iodide detector over the patient in a rectilinear fashion. The detected count rate modulated the intensity of a masked light bulb that scanned a film in an associated rectilinear pattern. While this device did produce images, it was very slow and had no capability for imaging rapidly changing distributions. The rectilinear scanner was used into the 1970s, but was finally supplanted by the Anger camera (2–5).

The Anger camera, also referred to as the scintillation camera (or gamma camera), is a radionuclide imaging device that was invented by Hal O. Anger. It is the predominant imaging system in nuclear medicine and is responsible for the growth and wide applicability of nuclear medicine. Anger was born in 1920. He received his BS degree in electric engineering from the University of California at Berkeley in 1943 and in 1946 he began working at the Donner Laboratories, where he developed a large number of innovative detectors and imaging devices including the scintillation well counter and a whole body rectilinear scanner using 10 individual sodium iodide probes. In 1957, he completed his first gamma

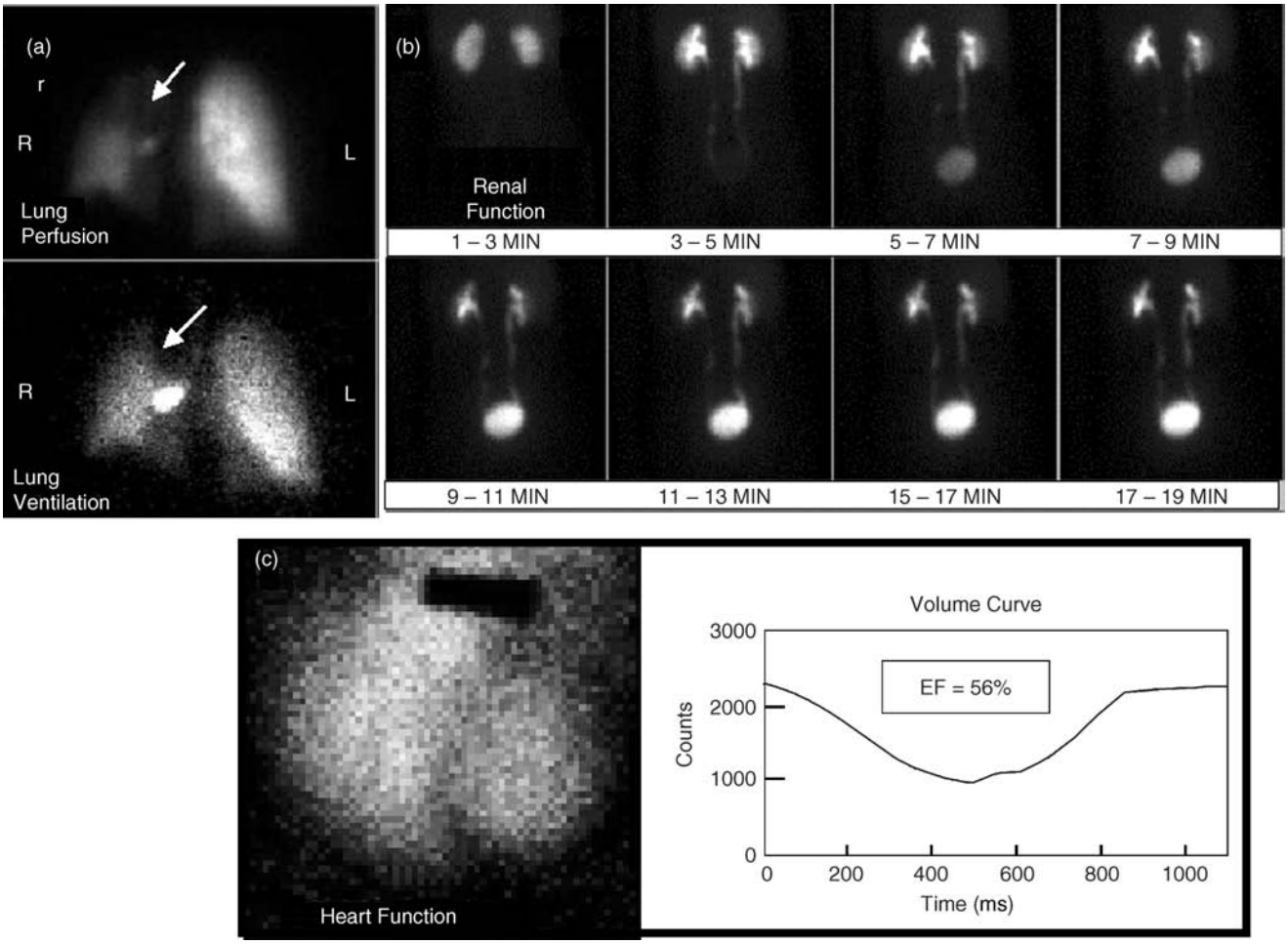


Figure 1. Anger camera clinical images. A. Lung ventilation and perfusion images are used to diagnose pulmonary emboli. B. Renal function images are used to diagnose a variety of problems including renal hypertension and obstruction. C. Gated blood pool studies permit evaluation of heart wall motion and ejection fraction.

imaging camera that he called a scintillation camera and is often referred to as the Anger camera (6). Anger's scintillation camera established the basic design that is still in use today. The first Anger camera had a 10 cm circular field of view, seven photomultiplier tubes, pinhole collimation, and could only be oriented in one direction. A picture of this initial scintillation camera is shown in Fig. 3 and a schematic drawing of the electronics is shown in Fig. 4. In 1959, Anger adapted the scintillation camera for imaging positron emitting radionuclides without collimation using coincidence between the camera and a sodium iodide detector. He also continued improving the scintillation camera for conventional gamma emitting radionuclides. By 1963, he had a system with a 28 cm field of view and 19 photomultiplier tubes (7). This device became commercialized as the nuclear Chicago scintillation camera. Throughout the 1960s, 1970s, and 1980s Anger remained active at Donner labs developing numerous other radionuclide imaging devices. He has received many prestigious awards including the John Scott Award (1964), a Guggenheim fellowship (1966), an honorary Doctor of Science degree from Ohio State University (1972), the Society of Nuclear Medicine Nuclear Pioneer Citation (1974), and the Society of Nuclear Medicine Benedict Cassen Award (1994) (8–10).

About the same time that the Anger camera was introduced, the molybdenum-99/technetium-99m radionuclide generator became available. This finding is mentioned because the advantages offered by this convenient source of ^{99m}Tc had a large influence on the development of the Anger camera. Technetium-99m emits a single gamma ray at 140 keV, has a 6 h half-life and can be attached to a large number of diagnostically useful compounds. Because it is available from a generator, it also has a long shelf life. The ^{99m}Tc is used in $> 80\%$ of nuclear medicine imaging studies. As a result, both the collimation and detector design of the Anger camera has been optimized to perform well at 140 keV (1).

SYSTEM DESCRIPTION

The Anger camera is a position sensitive gamma-ray imaging device with a large field of view. It uses one

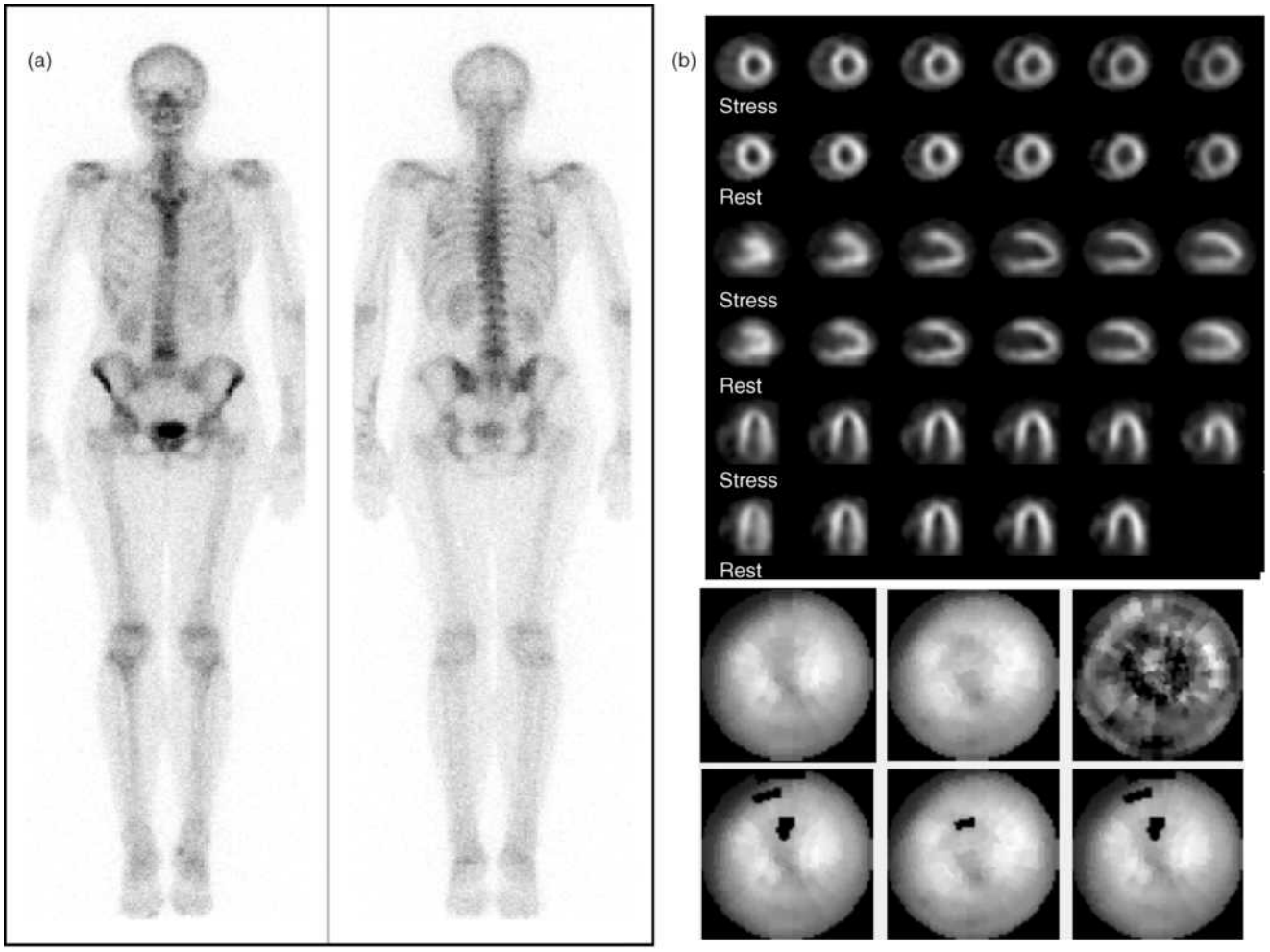


Figure 2. Anger camera clinical images. A. Bone scans are used to evaluate trauma and metastatic spread of cancer. B. Myocardial perfusion studies are used to evaluate coronary artery disease.

large, thin sodium iodide crystal for absorbing gamma-ray energy and converting that into visible light. The light signal is sampled by an array of photomultilier tubes that convert the light signal into an electronic pulse. The pulses from individual PMTs are combined in two ways. An energy pulse is derived from the simple summation of the PMT signals. The X and Y locations of the event are calculated from the sum of the PMT signals after position-dependent weighting factors have been applied. When a signal from a detected event falls within a preselected energy range, the X and Y locations are recorded in either list or frame modes. The components that make up the Anger camera are shown in Fig. 5 and are described in detail in the following section (11,12).

Sodium Iodide Crystal

Sodium iodide activated with thallium, NaI(Tl), is the detecting material used throughout nuclear medicine. Sodium iodide is a scintillator giving off visible light when it absorbs X- or gamma-ray energy. At room temperature, pure NaI has very low light emission, however, when small amounts (parts per million, ppm) of thallium are added, the

Figure 3. Initial Anger camera used to image a patient's thyroid with I-131. The field of view of this device was 10 cm. (Reprinted from Seminars in Nuclear Medicine, Vol 9, Powell MR, H.O. Anger and his work at Donner Laboratory, 164–168., 1979, with permission from Elsevier.)

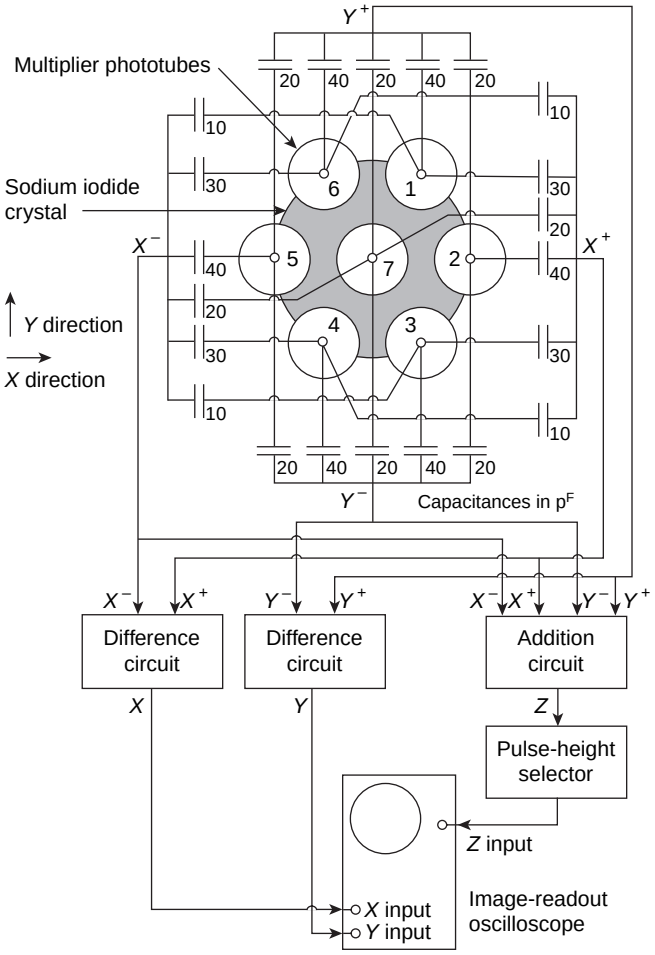


Figure 4. Electronic schematic for Anger's first scintillation camera. The photomultiplier tube position weighting was accomplished with a capacitor network. (From Instrumentation in Nuclear Medicine, Hine and Sorenson, Elsevier, 1967.)

efficiency for light emission is greatly enhanced. This is an especially important aspect for its application in the Anger camera since the light signal is used to determine both the energy and location of the gamma-ray interaction with the detector. In addition to its high light ouput, NaI(Tl) has several other desirable properties. It has a relatively high effective atomic number ($Z_{\text{eff}} = 50$) and the density is 3.67 g·cm^{-3}. This results in a high detection efficiency for gamma rays under 200 keV with relatively thin crystals (8,12–15).

Sodium iodide is grown as a crystal in large ingots at high temperatures (> 650 °C). The crystals are cut, polished, and trimmed to the required size. For Anger cameras, the crystals are typically 40 × 55 cm and 9.5 mm thick. Because NaI(Tl) absorbs moisture from the air (hygroscopic), it must be hermetically sealed. Any compromise of this seal often results in the yellowing of the crystal and its irreversible destruction.

In addition to the need to keep the crystal hermetically sealed, the temperature of the detector must be kept relatively constant. Temperature changes > 2 °C·h^{-1} will often shatter the detector.

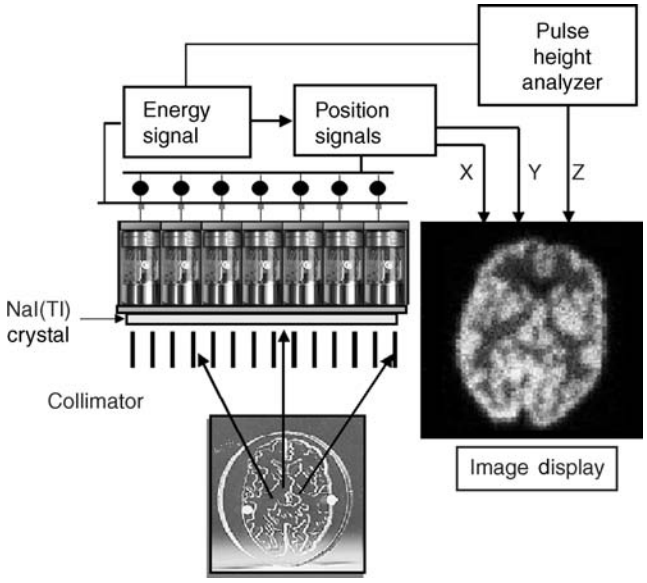


Figure 5. Anger camera components.

Light Pipe

The scintillation light generated in the crystal is turned into electronic signals by an array of photomultiplier tubes (PMTs). These signals provide both event energy and localization information. It is desirable that the magnitude of the signal from the photomultiplier tube be linearly related to the event location as shown in Fig. 6. However, when the PMTs are in close proximity to the crystal, the relationship between the signal magnitude and the event location is very nonlinear. In early designs of the Anger camera, a thick transparent material referred to as a light pipe was coupled to the crystals to improve spatial linearity and uniformity. Glass, lucite, and quartz have been used for this purpose. Design enhancement of the light pipe included sculptured contouring to improve light collection and scattering patterns at the PMT interface to reduce positional nonlinearities (Fig. 7). In the past decade, many of the spatial nonlinearities have been corrected algorithmically operating on digitized PMT signals. This has allowed manufacturers to either reduce the thickness of the light pipe or completely eliminate it (2,16,17).

PMT Array

The visible light generated by the absorption of a gamma ray in the NaI(Tl) crystal carries location and energy information. The intensity of the scintillation is directly proportional to the energy absorbed in the event. To use this information, the scintillation must be converted into an electronic signal. This is accomplished by photomultiplier tubes. In a PMT, the scintillation light liberates electrons at the photocathode and these electrons are amplified through a series of dynodes. The overall gain available from a photomultiplier tube is on the order 10^6.

Photomultiplier tubes are manufactured in a wide variety of shapes and sizes. Those with circular, hexagonal, and square photocathodes have all been used in Anger cameras. Hexagonal and square PMTs offer some advantages for

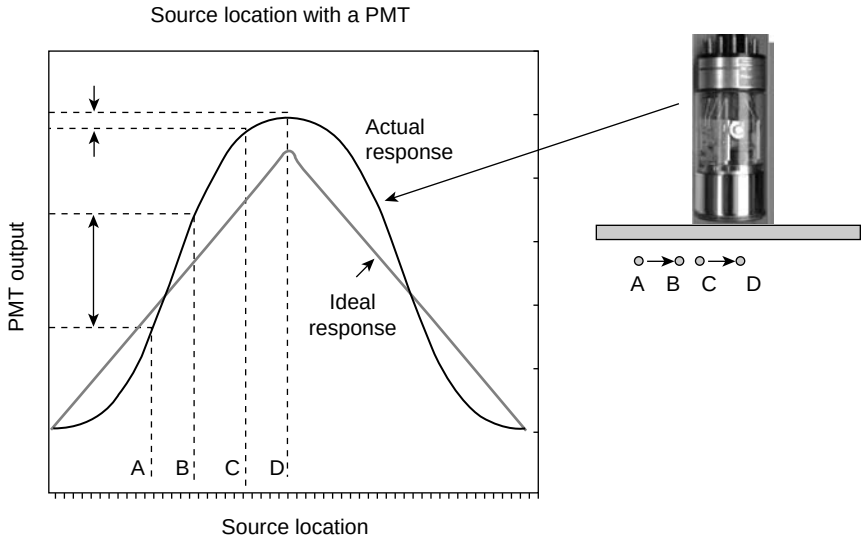


Figure 6. Photomultiplier tube response with source position. The ideal response can be approximated by interposing a light pipe between the crystal and photomultiplier tube.

close packing the PMTs over the surface of the detector. However, the sensitivity of all PMTs falls off near the edge of the field, so that "dead" space between the PMTs is unavoidable.

It is important to determine the energy of the detected event. Gamma rays that are totally absorbed produce a scintillation signal that is directly proportional to the gamma-ray energy. Thus, the signal resulting from the unweighted sum of the PMTs represents gamma-ray energy. This signal is sent to a pulse height analyzer. Scattered radiation from within the patient can be rejected by setting an energy window to select events that have resulted from the total absorption of the primary gamma ray. Gamma rays that have been scattered in the patient necessarily lose energy and are (largely) not included.

The position of the gamma-ray event on the detector is determined by summing weighted signals from the PMTs (2,6,7,12,16,18,19). Each PMT contributes to four signals: X^+, X^-, Y^+, Y^-. The magnitude of the contribution is determined both by the amount of light collected by the PMT and its weighting factor. For the tube located exactly at the center of the detector, the four weighting factors are equal. A tube located along the x axis on the left side (e.g., tube 5 in Fig. 4) contributes equally to Y^+ and Y^-, has a large contribution to X^-, and a small contribution to X^+. In Anger's original design, the weighting factors were provided by capacitors with different levels of capacitance (Fig. 4). In commercial units, the capacitor network was replaced by resistors (Fig. 8). In the past decades, the resistor weighting matrix has been largely supplanted by digital processing where nonlinear weighting factors can be assigned in software (Fig. 9).

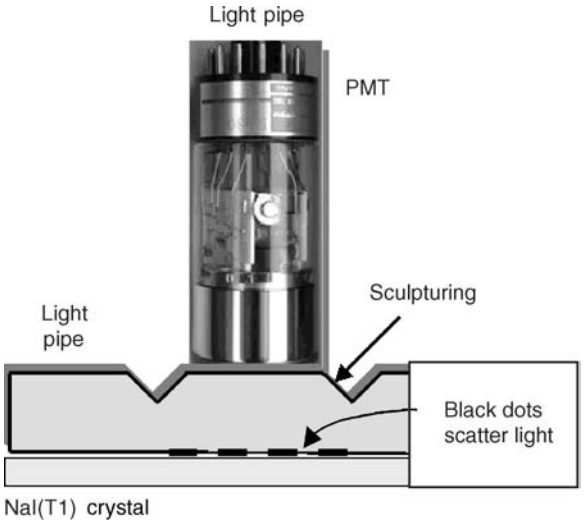


Figure 7. The light pipe is a transparent light conductor between the crystal and the photomultiplier tubes. The sculpturing grooves and black dot pattern spread the light to improve the positioning response of the photomultiplier tube.

It is clear that PMTs located near the event collect most of the scintillation light while those far away get relatively little. Because each PMT has an unavoidable noise component, the PMTs that receive the least light increase the error associated with the event localization. Initially, all the PMTs were included. Later, in order to eliminate PMTs that have little real signal, diodes were used to set current thresholds. In digital Anger cameras, the PMT thresholds are set in software (20–22).

The weighted signals from the PMTs are summed together to generate four position signals: X^+, X^-, Y^+, and Y^-. The X and Y locations are determined from: $(X^+ - X^-)/Z$ and $(Y^+ - Y^-)/Z$, where Z is the energy signal found from the unweighted sum of the PMT signals discussed above. This energy normalization is necessary to remove the size dependence associated with the intensity of the scintillation. This is not only important for imaging radionuclides with different gamma-ray energies, but it also improves the spatial resolution with a single gamma ray energy because of the finite energy resolution of the system. The energy signal is also sent to a pulse height analyzer where an energy window can be selected to include only those events associated with total absorption of the primary gamma.

Image Generation

Anger camera images are generated in the following way (see Fig. 10). A gamma ray is absorbed in the NaI(Tl)

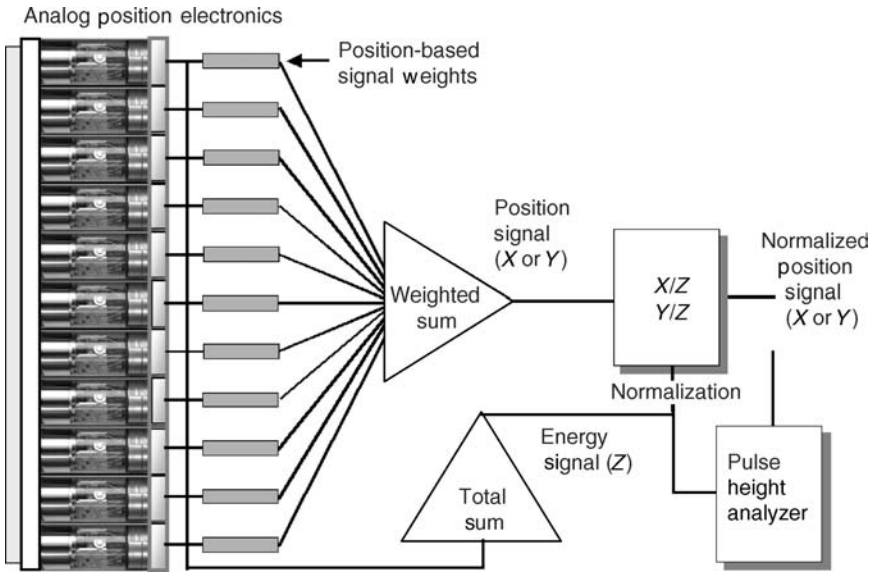


Figure 8. The Anger camera of the 1980s used resistors to provide position weighting factors. The energy signal was used both for normalization and scatter discrimination.

crystal and the resulting scintillation light is sampled by the PMT array to determine the event energy and location. The energy signal is sent to a pulse height analyzer and if the signal falls within the selected energy window, a logic pulse is generated. At the same time, the X and Y coordinates of the event are determined and the logic pulse from the PHA enables the processing of this information. For many years, Anger camera images were generated photographically with the enabled X and Y signals intensifying a dot on a cathode ray tube (CRT) viewed by a camera. In modern Anger cameras, the CRT has been replaced with computer memory and the location information is digital. The X and Y coordinate values, still enabled by the PHA, point to a memory element in a computer matrix. The contents of that memory element are incremented by 1. Information continues to accrue in the computer matrix until the count or time stopping criteria are met.

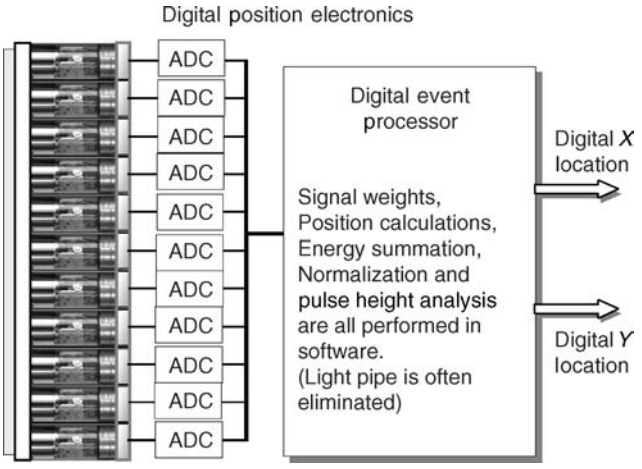


Figure 9. Digital Anger camera electronics. The photomultiplier tube signals are digitized so that signal weighting, energy and position determination are performed in software rather than with digital electronics.

The image generation described in the previous paragraph is referred to as frame or matrix mode. The information can also be stored in list mode where the X and Y coordinate of each event is stored sequentially along with a time marker. The list mode data can then be reconstructed at a later time to any desired time or spatial resolution.

Collimation

In order to produce an image of a radionuclide distribution, it has to be projected onto the detector. In a conventional camera, image projection is accomplished by the camera lens. However, gamma rays are too energetic to be focused with optics or other materials. The first solution to projecting gamma-ray images was the pinhole collimator. The pinhole collimator on an Anger camera is conceptually identical to a conventional pinhole camera. There is an inversion of the object and the image is magnified or minified depending on the ratio of the pinhole to detector distance and the object to pinhole distance. Pinhole collimators are typically constructed out of tungsten and require lead shielding around the "cone". Because the amount of magnification depends on the source to pinhole distance, pinhole images of large, three-dimensional (3D) distributions are often distorted. In addition, the count sensitivity falls off rapidly for off-axis activity. A better solution for most imaging situations is a multiholed parallel collimator (Fig. 11). As the name implies, the parallel collimator consists of a large number of holes with (typically) lead septae. Most parallel collimators have hexagonal holes that are $\sim$ 1.5 mm across and are 20–30 mm long. The septal walls are typically 0.2 mm thick. The parallel hole collimator produces projections with no magnification by brute force. Gamma rays whose trajectories go through the holes reach the detector while those with trajectories that intersect the septae are absorbed. Less than 1 out of 5000 gamma rays that hit the front surface of the collimator are transmitted through to the detector to form the image (2,7,11,12,20,23,24).

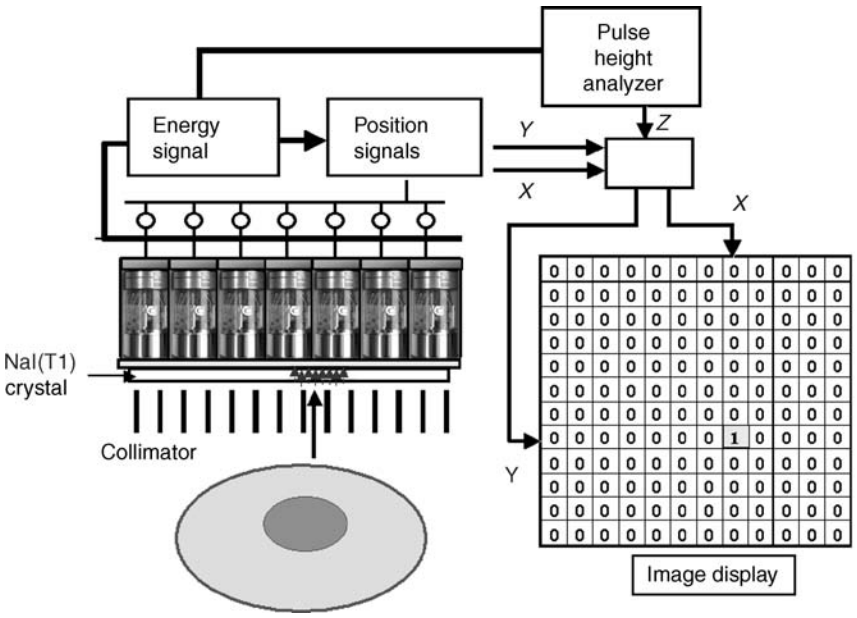


Figure 10. Schematic of Anger camera showing how image information is acquired.

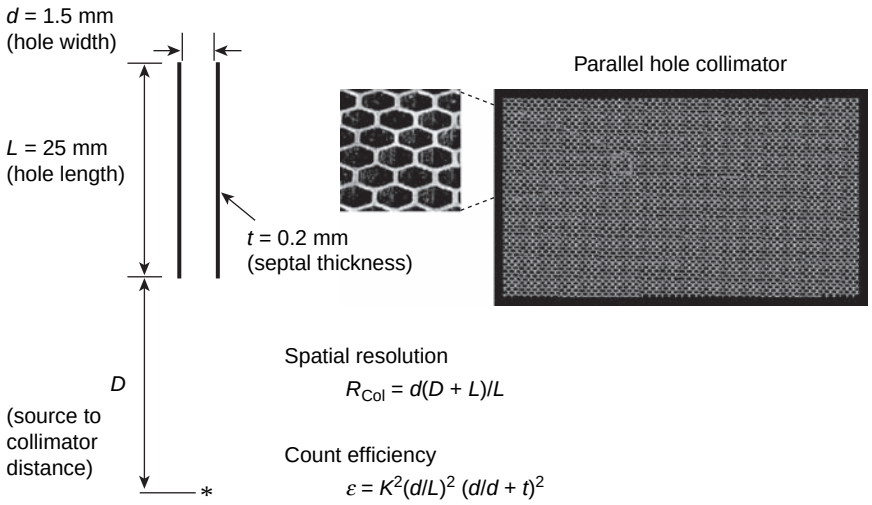


Figure 11. Collimation. Collimators are the image forming aperture of the Anger Camera, but are also the limiting component in spatial resolution and count sensitivity.

The spatial resolution of the collimator, R_{col}, is determined by the hole size (d), hole length (L), and the source to collimator distance (D): $R_{col} = d(L + D)/L$. The efficiency of a parallel hole collimator is expressed as $K^2(d/L)^2\ (d/d + t)^2$, where K is a shape factor constant equal to 0.26 and t is the septal wall thickness. Collimation design is an optimization problem since alterations in d and L to improve resolution will decrease count sensitivity. Collimator spatial resolution has a strong dependence on the source to collimator distance. As shown in Fig. 12, the spatial resolution rapidly falls with distance. However, the count sensitivity of a parallel hole collimator is *not* affected by the source distance because the parallel hole geometry removes the divergent rays that are associated with the inverse square loss. Another factor that influences collimator design is the energy of the gamma ray being imaged. Higher energy gamma rays require thicker septae and

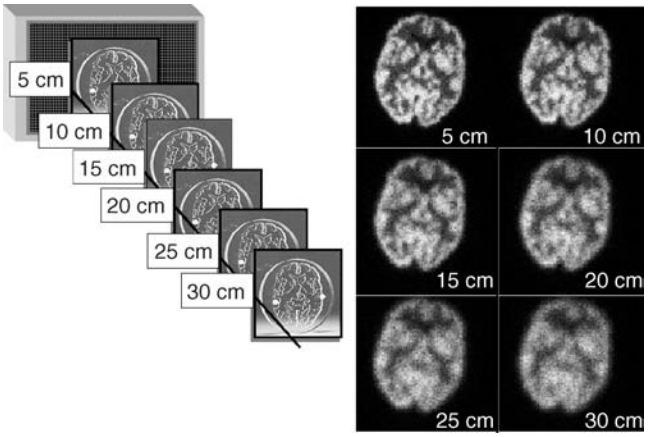


Figure 12. Spatial resolution dependence on source to collimator distance.

larger holes resulting in poorer resolution and count sensitivity (2,7,25).

CORRECTIONS

The analog Anger camera had a number of limitations because of nonlinearities in the position signals and because of the uneven light collection over the NaI(Tl) crystal. As digital approaches became viable over the last two decades, a number of corrections to improve energy, spatial, and temporal resolution have evolved. All of these corrections, and particular those involving spatial and energy resolution, require system stability. This challenge was significant because of the variation in PMT output associated with environmental conditions and aging. In order for corrections to be valid over an extended period of time, a method to insure PMT stability has to be implemented. Several different approaches have evolved. In one method, PMT gains are dynamically adjusted to maintain consistent output signals in response to stabilized light emitting diodes (LEDs). An LED is located beneath each PMT where its light is sampled 10–100 times·s^{-1}. Gains and offsets on the PMTs are adjusted so that the resulting signal is held close to its reference value. Another approach uses the ratio of photopeak/Compton plateau counts from a ^{99m}Tc or ^{57}Co source as the reference.

Flood Field Image

When the Anger camera is exposed to a uniform flux of gamma rays, the resulting image is called a flood field image. Flood field images are useful for identifying nonuniform count densities and may be acquired in two different ways. An intrinsic flood field is obtained by removing the collimation and placing a point source of activity 1.5–2 m from the detector. An extrinsic flood field is obtained with the collimator in place and with a large, distributed source (often called a flood source) placed directly on the collimator. Flood field sources using ^{57}Co are commercially available. Alternatively, water-filled flood phantoms are available into which ^{99m}Tc or other radionuclide can be injected and mixed.

Energy Correction

The energy signal represents the total energy absorbed in a gamma-ray interaction with the detector. This signal is determined by the intensity of the scintillation and by how much of the scintillation light is captured by the PMTs. Because the efficiency for sampling the scintillation is position dependent, there are fluctuations in the energy signals across the detector as shown in Fig. 13. These variations degrade energy resolution and have a significant effect on the performance of the scintillation camera that limit corrections for nonuniformity. The idea of using a reference flood field image to correct nonuniformities has been around for a long time. However, if the reference flood field image is acquired with little or no scattered radiation (as it often is), the correction factors are not appropriate during patient imaging. The reason is that scattered radiation and the amount of scatter entering the selected energy

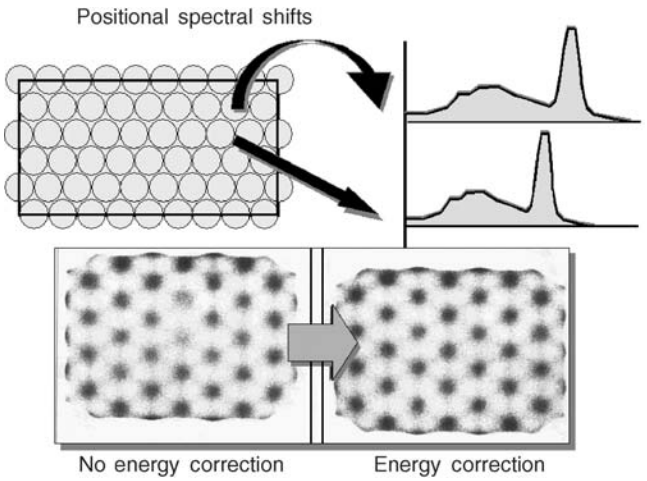


Figure 13. Energy correction. Local spectral gain shifts are evident across the crystal because of the variable sampling imposed by the photomultiplier tube array.

window will be position dependent. Energy correction electronics was introduced in the late 1970s that essentially generated an array of energy windows that are adjusted for the local energy spectra. Typically, the detector field of view is sampled in a 64 × 64 matrix and a unique energy window is determined for each matrix element. With the energy windows adjusted on the local photopeaks, the variations in the scatter component are greatly reduced. As shown in Fig. 13, energy correction does not significantly improve intrinsic field uniformity. Its role is to reduce the influence of source scatter on the other correction factors (11,20,26–28).

Spatial Linearity Correction

The nonlinearities in the PMT output with respect to source location causes a miss-positioning of events when Anger logic is used. This finding can be demonstrated by acquiring an image of a straight line distribution or a grid pattern. The line image will have a "wavy" appearance (Fig. 14). In the early 1980s, a method to improve the spatial linearity was developed. An imaging phantom array of precisely located holes in a sheet of lead is placed on the uncollimated detector and is exposed to a point source of ^{99m}Tc located 1–2 m away. The image of the hole pattern is used to calculate corrective x and y offsets for each point in the field of view. These correction factors are stored in a ROM. When an event is detected and the Anger logic produces x and y coordinates, the offsets associated with these coordinates are automatically added generating the new, accurate event location. Improving the spatial linearity has a profound affect on field uniformity as can be seen in Fig. 14 (17,28–31).

Uniformity Correction

After the energy and spatial linearity corrections have been made, there are still residual nonuniformities that are present in a flood field image. Typically, these will vary $< 10\%$ from the mean count value for the entire field. A high count reference flood field image can be acquired and

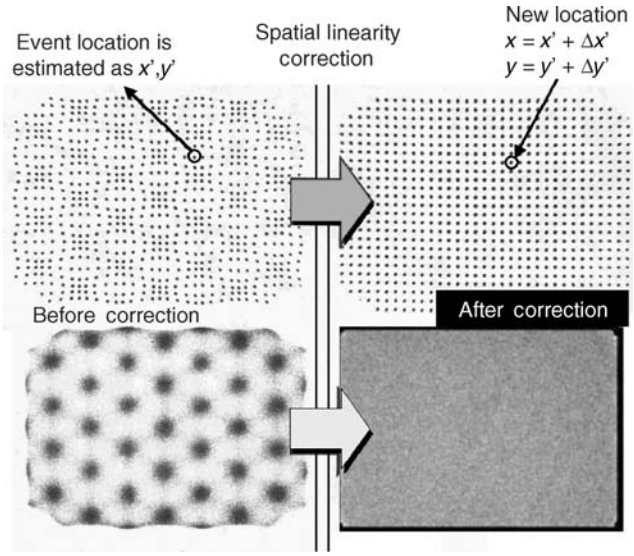


Figure 14. Spatial linearity correction. Accurate correction for inaccurate event localization has a profound effect on field uniformity.

this image is then used to generate regional flood correction factors that are then applied to subsequent acquisitions (Fig. 15) (17,29,32).

Pulse Pileup Correction

An Anger camera processes each detected event sequentially. Because the scintillation persists with a decay time of 230 ns, pulses from events occurring closely in time are distorted from summation of the light. This distortion is referred to as pulse pileup. As the count rate to the detector increases, the amount of pulse pileup also increases and becomes significant at count rates $> 30{,}000$ cps. For much of conventional nuclear medicine imaging, this is not a problem since the count rate is typically well below that level. There are certain applications such as coincidence positron emission tomography (PET) imaging where the detectors are exposed to event rates that can exceed 1,000,000 cps. Because the Anger logic used to establish the event location is essentially a centroid method, pulse pileup causes errors. An example of this is shown in Fig. 16, which shows an Anger camera image of four high-count rate sources. In addition to the actual sources, false images of source activity between the sources are also observed (33,34). The effects of pulse pileup can be minimized by electronic pulse clipping, where the pulse is forced to the baseline before all the light has been emitted and processed. While this increases the count rate capability, it compromises both spatial and energy resolution, which are optimal when the whole pulse can be sampled and integrated. One approach to reduce losses in spatial and energy resolution is to alter the integration time event-by-event, based on count rate demands. In addition, algorithms have been developed that can extrapolate the pulse to correct for

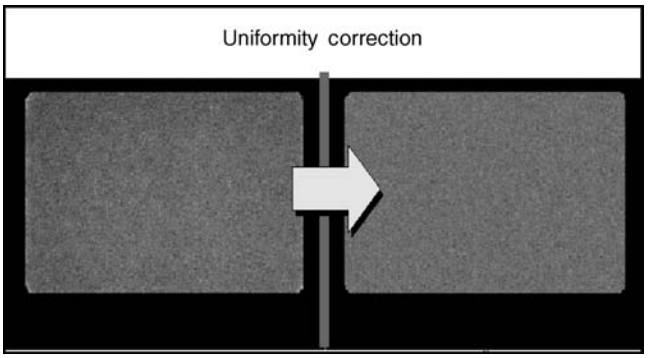


Figure 15. Uniformity correction. Residual non-uniformities can be reduced by skimming counts based on a reference flood field image.

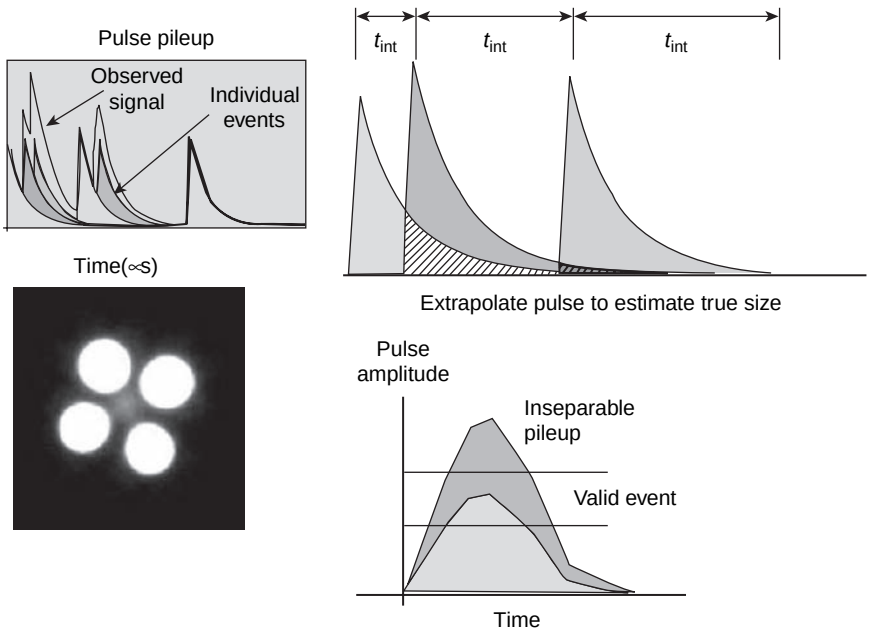


Figure 16. Pulse pileup correction. Pulse pileup correction improves the count rate capability and reduces the spurious placement of events.

its contribution to a second pileup pulse. This process can be repeated if a third pileup is also encountered. When this correction is performed at the PMT level, it reduces the "false" source images discussed above (35).

PERFORMANCE

Uniformity

When the Anger camera is exposed to a uniform flux of gamma rays, the image from that exposure should have a uniform count density. Anger cameras with energy, spatial linearity, and uniformity correction are uniform to within 2.5% of the mean counts.

Intrinsic Spatial Resolution

The intrinsic spatial resolution refers to the amount of blurring associated with the Anger camera independent of the collimation. It is quantified by measuring the full width at half-maximum (fwhm) of the line spread response function. The intrinsic spatial resolution for Anger cameras varies from 3 to 4.5 mm depending on the crystal thickness and the size of the PMTs. Another way of evaluating intrinsic spatial resolution for gamma-ray energies < 200 keV is with a quadrant bar phantom consisting of increasingly finer lead bar patterns where the bar width is equal to the bar spacing (Fig. 17). Typically an Anger camera can resolve a 2 mm bar pattern.

Extrinsic Spatial Resolution

The extrinsic spatial resolution, also referred to as the system spatial resolution, refers to the amount of blurring associated with Anger camera imaging. It depends on the collimation, gamma-ray energy, and the source to collimator distance. The standard method for determining the extrinsic resolution is from the fwhm of the line spread response function generated from the image of a line source positioned 10 cm from the collimator. Typical values for the extrinsic spatial resolution range from 8 to 12 mm.

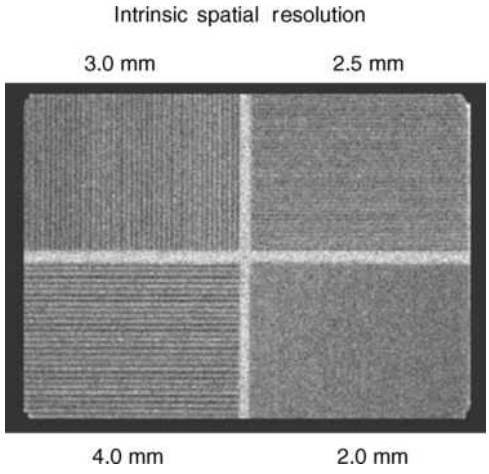


Figure 17. Intrinsic spatial resolution is routinely assessed with bar pattern images. An Anger camera can typically resolve the 2 mm bar pattern.

Energy Resolution

The energy signal generated from gamma-ray absorption in the detector has statistical fluctuations that broaden the apparent energy peaks of the gamma rays. Energy resolution is determined from $100\% \times \text{fwhm}/E\gamma$, where fwhm is of energy peak and $E\gamma$ is the gamma-ray energy. At 140 keV the energy resolution of an Anger camera is 10%. Good energy resolution is important because it permits the discrimination of scattered radiation from the patient. Gamma rays that are scattered in the patient necessarily loose energy and these scattered photons degrade image quality.

Spatial Linearity

Spatial linearity refers to the accurate positioning of detected events. On an Anger camera with spatial linearity correction, the misplacement of events is < 0.5 mm.

Multiwindow Spatial Registration

Because the Anger camera has energy resolution, it can acquire images from radionuclides that emit more than one gamma ray or from two radionuclides. However, the images from different energy gamma rays may have slightly different magnifications or have offsets because of imperfections in the energy normalization. The multiwindow spatial registration parameters quantifies the misalignment between different energy gamma rays (Fig. 18). For Anger cameras, the multiwindow spatial registration is < 2 mm, which is well below the system spatial resolution and therefore is not perceptible.

Count Rate Performance

The count rate capability of Anger camera ranges from 100,000 to 2,000,000 cps depending on the sophistication of the pulse handling technology as discussed above. Anger cameras that are used for conventional nuclear medicine imaging are designed to operate with maximum count rates of 200,000–400,000 cps range, whereas Anger cameras that are used for coincidence imaging require count rate capabilities that exceed 1,000,000 cps (25,28,30,32,36–41).

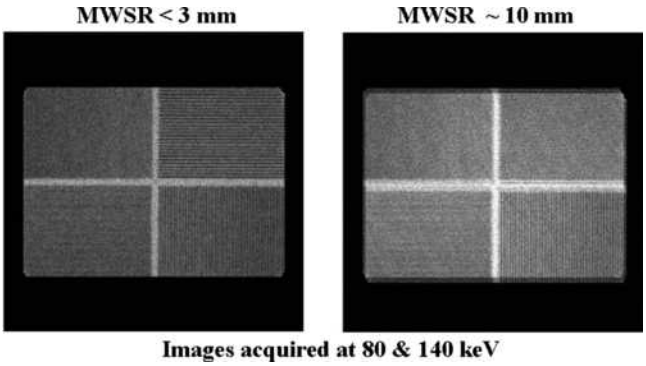


Figure 18. Multi-window spatial registration refers to ability to accurately image different gamma ray energies simultaneously. The figure on the right is an example of poor multi-window spatial registration.

SUMMARY

The Anger camera has been the primary imaging device in nuclear medicine for > 30 years and is likely to remain in that role for at least the next decade. Although it has evolved with the development of digital electronics, the basic design is essentially that promulgated by H.O. Anger. Special purpose imaging instruments based on semiconductor cadmium zinc telluride detectors are actively being pursued as imaging devices for ^{99m}Tc and other low energy gamma emitters. Their pixilated design removes the need for Anger logic position determination and the direct conversion of the absorbed energy into an electronic signal removes the need for photomultiplier tubes allowing compact packaging. However, over the range of gamma-ray energies encountered in nuclear medicine, NaI(Tl) still provides the best efficiency at a reasonable cost.

BIBLIOGRAPHY

Cited References

1. Banerjee S, Pillai MR, Ramamoorthy N. Evolution of Tc-99m in diagnostic radiopharmaceuticals. Semin Nucl Med 2001;31: 260–277.
2. Hine GJ, editor. Instrumentation in Nuclear Medicine. Volume 1, New York: Academic Press; 1967.
3. Pollycove M, Fish MB, Khentigan A. Clinical radioisotope organ imaging–diagnostic sensitivity and practical factors. Rectilinear scanner versus the anger-type scintillation camera. J Nucl Med 1967;8:321–322.
4. Blahd WH. Ben Cassen and the development of the rectilinear scanner. Semin Nucl Med 1996;26:165–170.
5. McCready VR. Milestones in nuclear medicine. Eur J Nucl Med 2000;27:S49–S79.
6. Anger H. A new instrument for mapping gamma ray emitters. Bio Med Quart Rep 1957; UCRL-3653 38.
7. Anger H. Scintillation camera with multichannel collimators. J Nucl Med 1964;5:515–531.
8. Hine GJ. The inception of photoelectric scintillation detection commemorated after three decades. J Nucl Med 1977;18:867–871.
9. Powell MR. H.O. Anger and his work at the Donner Laboratory. Semin Nucl Med 1979;9:164–168.
10. Tapscott E. Nuclear medicine pioneer: Hal O. Anger. First scintillation camera is foundation for modern imaging systems. J Nucl Med 1998;39:15N, 19N, 26N–27N.
11. Murphy PH, Burdine JA. Large-field-of-view (LFOV) scintillation cameras. Semin Nucl Med 1977;7:305–313.
12. Cherry S, Sorenson J, Phelps M. Physics in Nuclear Medicine. Philadelphia: W. B. Saunders; 2003.
13. Muehllehner G. Effect of crystal thickness on scintillation camera performance. J Nucl Med 1979;20:992–993.
14. Royal HD, Brown PH, Claunch BC. Effects of a reduction in crystal thickness on Anger-camera performance. J Nucl Med 1979;20:977–980.
15. Keszthelyi-Landori S. NaI(Tl) camera crystals: imaging capabilities of hydrated regions on the crystal surface. Radiology 1986;158:823–826.
16. Anger H. Scintillation Camera. Rev Sci Instrum 1958;29: 27–33.
17. Genna S, Pang SC, Smith A. Digital scintigraphy: principles, design, and performance. J Nucl Med 1981;22:365–371.
18. Anger H. Scintllation camera with 11 inch crystal. UCRL-11184 (1963).
19. Scrimger JW, Baker RG. Investigation of light distribution from scintillations in a gamma camera crystal. Phys Med Biol 1967;12:101–103.
20. White W. Resolution, sensitivity, and contrast in gamma-camera design: a critical review. Radiology 1979;132:179–187.
21. Zimmerman RE. Gamma cameras—state of the art. Med Instrum 1979;13:161–164.
22. Goodwin PN. Recent developments in instrumentation for emission computed tomography. Semin Nucl Med 1980;10:322–334.
23. Strand SE, Lamm IL. Theoretical studies of image artifacts and counting losses for different photon fluence rates and pulse-height distributions in single-crystal NaI(T1) scintillation cameras. J Nucl Med 1980;21:264–275.
24. Kereiakes JG. The history and development of medical physics instrumentation: nuclear medicine. Med Phys 1987;14:146–155.
25. Chang W, Li SQ, Williams JJ, Bruch PM, Wesolowski CA, Ehrhardt JC, Kirchner PT. New methods of examining gamma camera collimators. J Nucl Med 1988;29:676–683.
26. Budinger TF. Instrumentation trends in nuclear medicine. Semin Nucl Med 1977;7:285–297.
27. Myers WG. The Anger scintillation camera becomes of age. J Nucl Med 1979;20:565–567.
28. Heller SL, Goodwin PN. SPECT instrumentation: performance, lesion detection, and recent innovations. Semin Nucl Med 1987;17:184–199.
29. Muehllehner G, Colsher JG, Stoub EW. Correction for field nonuniformity in scintillation cameras through removal of spatial distortion. J Nucl Med 1980;21:771–776.
30. Muehllehner G, Wake RH, Sano R. Standards for performance measurements in scintillation cameras. J Nucl Med 1981;22: 72–77.
31. Johnson TK, Nelson C, Kirch DL. A new method for the correction of gamma camera nonuniformity due to spatial distortion. Phys Med Biol 1996;41:2179–2188.
32. Murphy PH. Acceptance testing and quality control of gamma cameras, including SPECT. J Nucl Med 1987;28:1221–1227.
33. Strand SE, Larsson I. Image artifacts at high photon fluence rates in single-crystal NaI(T1) scintillation cameras. J Nucl Med 1978;19:407–413.
34. Patton JA. Instrumentation for coincidence imaging with multihead scintillation cameras. Semin Nucl Med 2000;30: 239–254.
35. Wong WH, Li H, Uribe J, Baghaei H, Wang Y, Yokoyama S. Feasibility of a high-speed gamma-camera design using the high-yield-pileup-event-recovery method. J Nucl Med 2001;42: 624–632.
36. O'Connor MK, Oswald WM. The line resolution pattern: a new intrinsic resolution test pattern for nuclear medicine [see comments]. J Nucl Med 1988;29:1856–1859.
37. De Agostini A, Moretti R. Gamma-camera quality control procedures: an on-line routine. J Nucl Med Allied Sci 1989;33:389–395.
38. Lewellen TK, Bice AN, Pollard KR, Zhu JB, Plunkett ME. Evaluation of a clinical scintillation camera with pulse tail extrapolation electronics. J Nucl Med 1989;30:1554–1558.
39. Links JM. Toward a useful measure of flood-field uniformity: can the beauty in the eye of the beholder be quantified? [editorial] [see comments]. Eur J Nucl Med 1992;19:757–758.
40. Hander TA, Lancaster JL, Kopp DT, Lasher JC, Blumhardt R, Fox PT. Rapid objective measurement of gamma camera resolution using statistical moments. Med Phys 1997;24:327–334.
41. Smith EM. Scintillation camera quality control, Part I: Establishing the quality control program. J Nucl Med Technol 1998;26:9–13.

See also COMPUTED TOMOGRAPHY, SINGLE PHOTON EMISSION; IMAGING DEVICES; MICROPOWER FOR MEDICAL APPLICATIONS; NUCLEAR MEDICINE INSTRUMENTATION.

ANGIOPLASTY. See CORONARY ANGIOPLASTY AND GUIDEWIRE DIAGNOSTICS.

ANORECTAL MANOMETRY

ASHOK K. TUTEJA
University of Utah
Salt Lake City, Utah

GEORGE E. WAHLEN
Veterans Affairs Medical Center
and the University of Utah
Salt Lake City, Utah

SATISH S.C. RAO
University of Iowa College of
Medicine
Iowa City, Iowa

INTRODUCTION

The most commonly performed test is the evoluation of anorectal function. These tests can provide useful information regarding the pathophysiology of disorders that affect defecation, continence, or anorectal pain. Anorectal manometry quantifies anal sphincter tone and assesses anorectal sensory response, recto anal reflexes, and rectal compliance. Sensory testing is usually performed along with anorectal manometry and is generally considered a part of the manometry.

The functional anatomy of the anorectum, the equipment, and the technique used for performing anorectal manometry and the parameters for measuring and interpreting the test are discribed in this article. The indications for anorectal manometry are shown in Table 1.

FUNCTIONAL ANATOMY AND PHYSIOLOGY OF THE ANORECTUM

The neuromuscular integrity of the rectum, anus, and the pelvic floor musculature help to maintain normal fecal continence and evacuation. The rectum is an S-shaped muscular tube, which serves as a reservoir and as a pump for emptying stool. The anus is a 2–4 cm long muscular cylinder, which at rest forms an angle with the axis of the rectum of $\sim 90^\circ$. During voluntary squeeze the angle becomes more acute, $\sim 70^\circ$, and during defecation, the angle becomes more obtuse, ~ 110–130° (1,2). The puborectalis muscle, one of the pelvic floor muscles, is responsible for these changes. The anal canal is surrounded by specialized muscles that form the anal sphincters [internal anal sphincter (IAS) and the external anal sphincter (EAS)]. The IAS is 0.5 cm thick and is an expansion of circular smooth muscle layer of the colon. It is an involuntary muscle innervated by fibers of the autonomic nervous system. The EAS is composed of striated muscle, is 0.6–1 cm thick, and is under voluntary control (3). The anus is normally closed by the tonic activity of the IAS. This barrier is reinforced during voluntary squeeze by the EAS. The IAS contributes ~ 70–85% of the resting anal pressure. The IAS does not completely seal the anal canal and requires the endo-anal cushions to interlock and seal the canal. The anal mucosal folds, together with the expansive anal vascular cushions, provide a tight seal. These barriers are further augmented by the puborectalis muscle, which forms a flap-like valve that creates a forward pull and reinforces the anorectal angle to prevent fecal incontinence (3,4). The rectum and the IAS are innervated by the autonomic nervous system. The EAS and the anoderm are supplied by somatic nerves. The mucosa of the rectum and proximal anal canal is lack of somatic sensory innervation. The pudendal nerve, arising from second, third, and fourth sacral nerves is the principal somatic nerve and innervates the EAS, the puborectalis muscle, and the anal mucosa (5).

Table 1. Indications for Anorectal Manometry

Fecal Incontinence
Chronic idiopathic constipation
Diagnosis of Hirschsprung's disease and/or follow up
Megarectum
Pelvic floor dyssynergia
Rectocele
Solitary rectal ulcer
Rectal prolapse
Functional anorectal pain
Neurological diagnostic investigations
Biofeedback training
Pre- and Postsurgery (pouch)

EQUIPMENT FOR ANORECTAL MANOMETRY

The manometric system has two major components: the manometric probe and the pressure recording apparatus. Several types of probes and pressure recorders are available. Each system has distinct advantages and disadvantages. The most commonly used probes and recording devices are reviewed here (6).

Water-Perfused Catheter

This catheter has multiple canals through which water is perfused slowly using a pneumohydraulic pump (Arndorfer, Milwaukee, WI; MUI Scientific Ltd., Toronto, Canada). The infusion rate is 0.5 mL $\cdot$ canal^{-1} $\cdot$ min^{-1}. In the catheter with helicoidal configuration the side holes of the canals are arranged radially and spaced 1, 2, 3, 4, 5, and 8 cm from the "0" reference point. A compliant balloon is tied to one end of the probe, which has a 200 mL capacity. The catheter is placed inside the anorectum, but the pressure transducers are located outside the body and across the flow of water. Resistance generated to the flow of water by luminal contractile activity is quantified as intraluminal pressure. The transducers located on the perfusion pump and the perfusion ports must be at the same level during calibration and when performing the study. The maintenance of the water perfused system requires relatively skilled personnel and air bubbles in the water tubing can affect the pressure recordings. The probe and the recording system are inexpensive and versatile. The closely spaced pressure sensors along the probe can record rectal and anal canal pressures and discriminate between EAS and IAS activity (7).

Solid-State Probe

This system has pressure sensors (microtransducers) that are mounted on the probe. This allows more accurate measurement by placing the pressure sensors directly at the source of pressure activity. The transducers are true strain gauge, that is, they consist of a pressure sensitive diaphragm with semiconductor strain gauges that are mounted on its inner surface. Currently, this is the most accurate catheter system for performing manometery (8). It is user friendly, offers higher fidelity, and is free of limitations imposed by the perfused system. However, it is expensive.

Amplifier–Recorder

The pressure signals that are obtained from the transducer are amplified and recorded on computerized small size amplifiers and recorders (e.g., Polygraph-Medronics/Functional Diagnostics, Minneapolis, MN; Insight, Sandhill Scientific Ltd. Littleton, CO; 7-MPR, Gaeltec, Isle of Sky, UK, and others). They are small, compact, and not only serve as amplifiers and recorders, but also facilitate analysis of data and provide convenient storage for future retrieval of data or for generating a database. No one system is ideal, although each has its strengths and weakness.

STUDY PROTOCOL

General Instructions for Patients Undergoing Anorectal Manometry

In order to maximize uniformity, the manometry should be accomplished with the rectum emptied of feces. The preparation cannot be indispensable for incontinent patients. Constipated patients must be examined several hours after a 500 mL tap water enema or a single Fleets phospho-soda enema. Patients may continue with their routine medications, but the medications should be documented to facilitate interpretation of the data. Patients may eat or drink normally up to the time of the test. Upon arrival at the motility laboratory, the patient may be asked to change into a hospital gown.

The duration of the test is $\sim$ 1 h. The manometry catheter is inserted into the rectum while patients lie on their left side. Patients will feel movement of the catheter and distension of the balloon. After the test, patients can drive home and resume their usual work and diet. It is a safe procedure. There should be little, if any, discomfort during manometry. No anesthetic is used. Absolute contraindication to manometry is recent surgery of the rectum and anal canal, relative contraindication is a poorly compliant patient and rectum loaded with stool.

Patient Position and Digital Examination

It is recommended that the patient is placed in the left lateral position with knees and hips bent to 90°. After explaining the procedure, a digital rectal examination is performed using a lubricated gloved finger. The presence of tenderness, stool, or blood on the finger glove should be noted.

Probe Placement

Next, the lubricated manometry probe is gently inserted into the rectum and oriented such that the most distal sensor (1 cm level) is located posteriorly at 1 cm from anal verge. The markings on the shaft of the probe should aid this orientation.

Run-in Time

After probe placement, a rest (run-in) period should be allowed ($\sim$ 5 min) to give the subject time to relax and allow the sphincter tone to return to basal levels.

Resting Anal Pressure

Currently, two methods are available for assessing this function (9). *Station pull-through*: In this technique, the most distal sensor of a multiport catheter assembly is placed 5 cm above the anal margin. At every 30 s intervals, the catheter is withdrawn by 0.5 cm either manually or with a probe withdrawal device (10). As the sensors straddle the high pressure zone, there is a step up of pressure. The length and the highest pressure of the anal sphincter is then measured. Because pull-through excites anal contraction and the individual is conscious of these movements, the recorded pressure is high (10). For the same reason, a rapid pull through is not an accurate method and is not advisable for measuring anal sphincter function. *Stationary method*: Uses radially arranged multiport catheter, at least three sensors, 1 cm apart that is placed in the anal sphincter zone, that is, 0–3 cm from the anal verge (11). After allowing the tracings to stabilize, the highest sphincter pressure that is observed at any level in the anal canal is taken as the maximum resting sphincter pressure. Resting pressures can be expressed as the average obtained from each transducer or as a range to identify asymmetry of anal canal pressures (12).

Normal anal canal pressures vary according to sex, age, and techniques used (10). Normal values for anorectal manometry are shown in Table 2. There are normal variations in external sphincter pressures both radially and longitudinally (12,14). Anterior quadrant pressures are lower in the orad part of anal canal while posterior quadrant pressures are lower in the distal part of the anal canal. In the mid-anal canal, pressures are equal in all four quadrants. Manometry also enables routine calculation

Table 2. Suggested List of Tests/Maneuver Based on Indication (s)[a]

	Indications for maneuver	
Test	Incontinence	Constipation
Resting pressure	Yes	Yes
Squeeze pressure/duration	Yes	No
Cough reflex	Yes	No
Attempted defecation	No	Yes
RAIR	No	Yes
Rectal sensation	Yes	Yes
Rectal compliance	Optional	Optional

[a]From Ref. 13 with permission.

of anal canal length. Overall pressures are higher in men and younger persons and men have longer anal canals than women. But there is considerable overlap in values and disagreement among various studies about the effect of age and gender on anal canal pressures (10–12,15). Furthermore, subjects with values outside the normal range may not have clinical symptoms and patients with clinical symptoms may exhibit normal values (16).

Squeeze Sphincter Pressure

This pressure can be measured with either the station pull-through or the alternative technique. In the station pull-through technique; after placing the multiport assembly as describe above, at each level the subject is asked to squeeze and to maintain the squeeze for as long as possible (at least 30 s). Alternatively, with a multiport catheter is place, the subject is instructed to squeeze on three separate occasions, with a minutes' rest between each squeeze to allow for recovery from fatigue. The average of the three highest sphincter pressures recorded at any level in the anal canal is taken as the maximum anal squeeze pressure (13). The duration of maximum sustained squeeze should also be determined and is defined as the time interval in seconds during which the subject can maintain a squeeze pressure at or above 50% of the maximum pressure.

Weak squeeze pressures may be a sign of external sphincter damage, neurological damage of the motor pathways, or a poorly compliant patient. Squeeze pressures should be evaluated together with response to cough reflex (16).

Response to Increases in Intraabdominal Pressure

An increase in intraabdominal pressure brought about by asking the subject to blow up a party balloon or by coughing is associated with a reflex increase in the activity of the EAS (11); also called the cough reflex. This reflex response causes the anal sphincter pressure to rise above that of the rectal pressure so that continence is maintained. The response may be triggered by receptors in the pelvic floor and mediated through a spinal reflex arc. In patients with complete supra conal spinal cord lesions, this reflex response is present, but the voluntary squeeze may be absent whereas in patients with lesions of the cauda equina or of the sacral plexus, both the reflex response and the voluntary squeeze are absent.

Rectoanal Pressure Changes During Attempted Defecation

In this maneuver, the subject is asked to bear down, and simulate the act of defecation. The side holes of catheter are located within the anal canal and the rectal balloon is kept inflated. The normal response consists of an increase in rectal pressure coordinated with a relaxation of the intraanal pressure. Alternatively, there may be a paradoxical increase in anal canal pressures, or absent relaxation or incomplete relaxation of the anal sphincter (Fig. 1) (17). It must be appreciated that laboratory conditions may induce artifactual changes, which is a learned response and is under voluntary control.

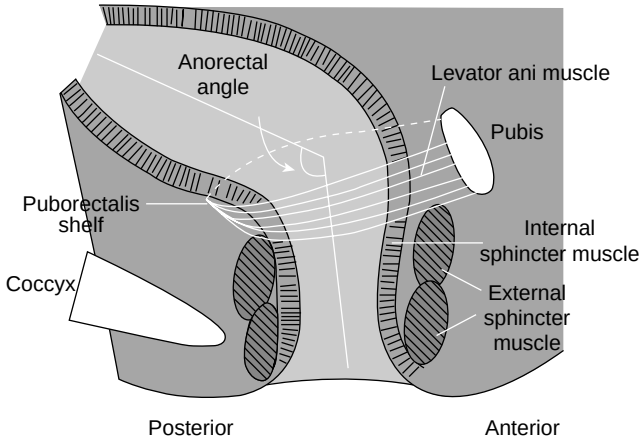


Figure 1. Structures of the anorectum: Reprinted from Ref. 17 with permission from American Gastroenterological Association.

Rectoanal Inhibitory Reflex

This consists of reflex relaxation of the IAS in response to rectal distension. The catheter is positioned with its side holes within the anal canal. Volumes of air are rapidly inflated in the rectal balloon and removed. The inflated time is 10 mL · s. The reflux is evoked with 10, 20, 40, 60, 80, 140, and 200 mL. As the volume of rectal distension is increased, the amplitude and duration of IAS relaxation increases (7). The absolute or relative amplitude of the IAS relaxation depends on the preexisting tone of the IAS and the magnitude of its contribution to the basal anal tone. This reflex may facilitate sampling of rectal contents by the sensory receptors in the upper anal canal and may also help to discriminate flatus, from liquid or solid stools. This reflex is regulated by the intrinsic myenteric plexus. In patients with Hirschsprung's disease and in those with a history of rectal resection and colo- or ileo-anal anastomosis, this reflex is absent. However, in patients with spinal cord injury and in patients with transaction of the hypogastric nerves or lesions of the sacral spinal cord, it is present (18).

Sensory Testing

Rectal Sensory Function. In this technique, the rectal sensory threshold for three common sensations (first detectable sensation, the sensation of urgency to defecate, and the sensation of pain or maximum tolerable volume) is assessed. This can be assessed either by the intermittent rectal distension or by the ramp inflation method.

Intermittent Rectal Distension. This technique is performed by inflating a balloon in the rectum using a handheld syringe. After each inflation, the balloon is deflated completely and after a rest period it is reinflated to the next volume (19).

Ramp Inflation. In this method, the rectum is progressively distended without deflation. This is performed by continuously inflating the balloon at a constant rate with a peristaltic pump or a syringe using increasing volumes of air or fluid or in a stepwise fashion, with a 1 min interval between each incremental inflation of 10–30 cm^3. It is known that the type of inflation (phasic vs. continuous) and the speed of continuous inflation affect the threshold

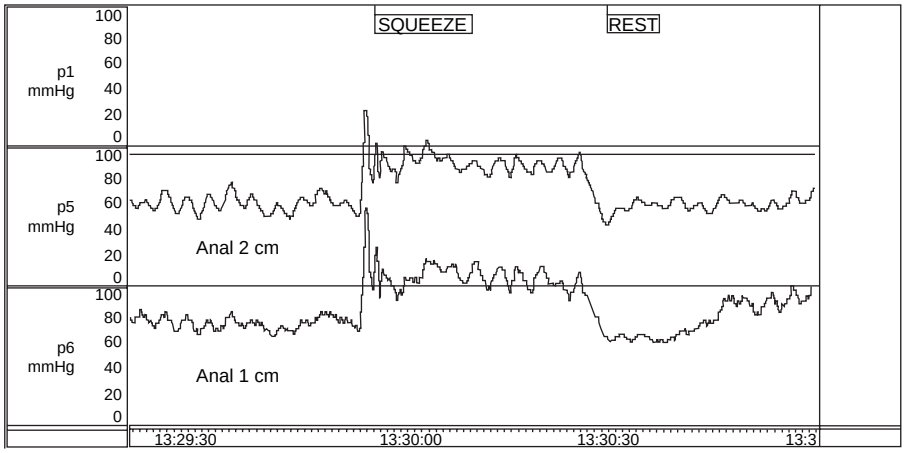


Figure 2. Normal squeeze profile.

volume required for healthy control subjects to perceive distension (20). Also the size and shape of the balloon will affect the threshold volume. Some of this variability can be reduced by using a high compliance balloon and a continuous-infusion pump or a barostat (21).

The maximum tolerable volume or pain threshold may be reduced in patients with a noncompliant rectum (e.g., proctitis) abdominoperineal pull-through, and rectal ischemia (9). Pain threshold also may be reduced in patients with irritable bowel syndrome (22). Higher sensory threshold is seen in autonomic neuropathy, congenital neurogenic anorectal malformations (spinal bifida, Hirschprung's disease, meningocele) and with somatic alteration in rectal reservoir (megarectum, descending perineum syndrome) (20,23). Rectal sensory threshold is altered by change in rectal wall compliance and sensory data should be interpreted along with measurement of rectal compliance (24).

Anal Sensation. At present, assessment of anal canal sensation is not of established value for the diagnosis and treatment of patients with constipation or fecal incontinence (9).

Rectal Compliance

The capacity and distensibility of the rectum are reflected by its compliance. It is a measure of the rectal reservoir function and is defined as the change in rectal volume per unit change in rectal pressure (11). The rectal compliance can be measured by the balloon distension method or more accurately by using a computerized barostat. The higher the compliance, the lower the resistance to distension and vice versa. Low rectal compliance is also seen in patients with acute ulcerative colitis, radiation proctitis, and low spinal cord lesions (20). High compliance is seen in patients with megarectum. Decreased rectal compliance can result in decreased rectal capacity, fecal urgency, and may contribute to fecal incontinence (25).

MANOMETRIC FEATURES OF FECAL INCONTINENCE AND CONSTIPATION

Fecal Incontinence

Anorectal manometry can provide useful information regarding the pathophysiology and management of fecal incontinence (26). Anal sphincter pressures may be decreased in patients with fecal incontinence; either circumferentially or in one quadrant of the anal canal (Fig. 2). Manometry can also determine if compensatory squeeze pressure can be activated. A reduced resting pressure correlate with predominant weakness of IAS and decreased squeeze pressures correlate with EAS defects (27). Two large studies have reported that maximum squeeze pressure has the greatest sensitivity and specificity in discriminating fecal incontinence from continent and healthy controls (28,29). The ability of the EAS to contract reflexly can also be assessed during abrupt increases of intra-abdominal pressure (e.g., when coughing). This reflex response causes the anal sphincter pressure to rise above that of the intrarectal pressure to preserve continence. This reflex response is absent in patients with lesions of the cauda equina or the sacral plexus (18,30). On sensory testing, both hyper- and hyposensitivity can be seen. Assessment of rectal sensation is useful in patients with fecal incontinence associated with neurogenic problems, such as diabetes mellitus (decrease in rectal sensations) or multiple sclerosis (increase in rectal sensation) (31). In some patients, rectal sensory thresholds may be altered because of changes in the compliance of the rectal wall. Patients with megarectum have decreased rectal sensation; and can present with fecal incontinence. Patients with incontinence often have lower rectal compliance (i.e., chronic rectal inschemia, proctitis).

Because of the wide range of normal values in anorectal physiologic testing, no single test can predict fecal incontinence. However, a combination of the tests with clinical evaluation is helpful in assessment of patients with fecal incontinence (32). Anorectal manometry is also useful in evaluating the responses to biofeedback training as well as assessing objective improvement following drug therapy or surgery.

Constipation

Anorectal manometry is useful in the diagnosis of dyssynergic defecation. Manometry helps to detect abnormalities during attempted defecation. Normally, when subjects bear down or attempt to defecate, there is a rise in rectal pressure, which is synchronized with a relaxation of the EAS (Fig. 3). The inability to perform this coordinated

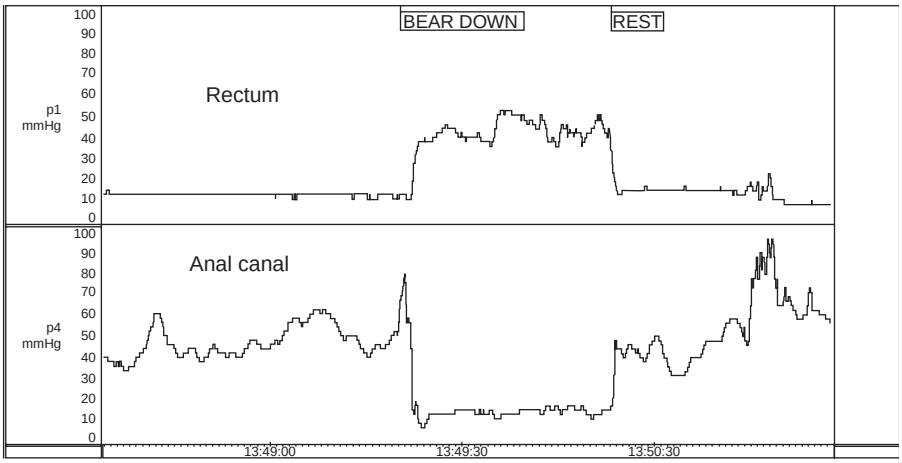


Figure 3. Strain maneuver: A normal coordinated response of the anorectum during attempted defecation shows a rise in rectal pressure associated with a decrease in anal sphincter pressure.

movement represents the chief pathophysiologic abnormality in patients with dyssynergic defecation (17). This inability may be due to impaired rectal contraction, paradoxical anal contraction, impaired anal relaxation, or a combination of these mechanisms (Fig. 4) (Fig. 5). Anorectal manometry also helps to exclude the possibility of Hirschsprung's disease. The absence of the rectoanal inhibitory reflex accompanied by a normal intrarectal pressure increase during distension of the intrarectal balloon is evidence of denervation of the intrinsic plexus at the recto-anal level. Megarectum can cause a falsely negative reflex. In this condition, there is hypotonia of the rectal wall due to a deficiency of visceroelastic properties of the rectum and high degrees of rectal distension are necessary to produce the reflex. In addition to the motor abnormalities, sensory dysfunction may be present. The rectal sensations are reduced in patients with megarectum. The first sensation and the threshold for a desire to defecate may be higher in $\sim$ 60% of patients with dyssynergic defecation (33). The threshold for urge to defecate may be absent on elevated in patients with chronic constipation. Maximum tolerable volume can also be elevated (34). But is not clear whether these findings are the cause or secondary to constipation. When rectal sensation is impaired, neuromuscular conditioning using biofeedback technique can be effective in improving the dysfunction.

Select Appropriate Test/Maneuver

Because anorectal manometry consists of several maneuvers, it is important to determine whether a patient needs all of the maneuvers or only a selection from the array of tests described below. The patient's symptoms and the reason for referral are helpful in choosing the appropriate list. A suggested list is given in Table 3.

Prolonged Anorectal Manometry. It is now feasible to perform anorectal manometry for prolonged periods of time outside the laboratory setting. With the use of this technique, it is possible to measure physiologic functions of the anal sphincter while the person is mobile and free (35). This technique shows promise as an investigational procedure, but its clinical applicability has not been established.

Clinical Utility and Problems with Anorectal Manometry. A systematic and careful appraisal of anorectal function can provide valuable information that can guide treatment of patients with anorectal disorders. Prospective studies have shown that manometric tests of anorectal function provide not only an objective diagnosis, but also a better understanding of the underlying pathophysiology. In

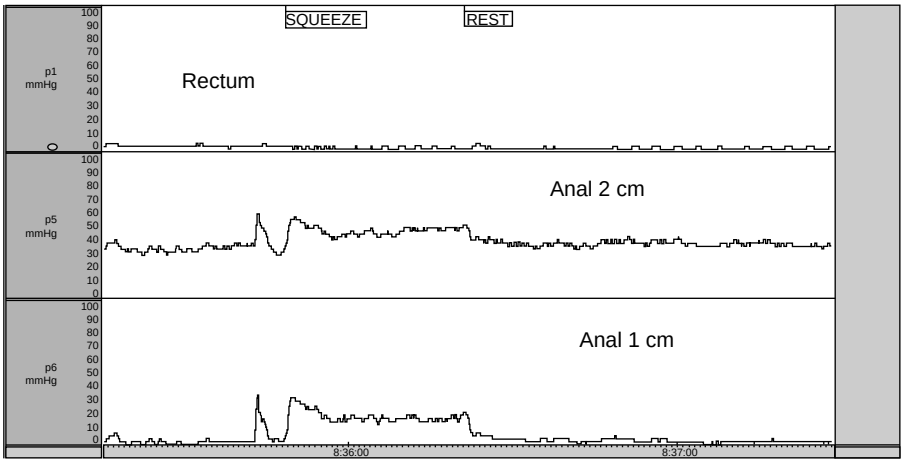


Figure 4. Weak resting and squeeze anal sphincter pressure in a patient with fecal incontinence.

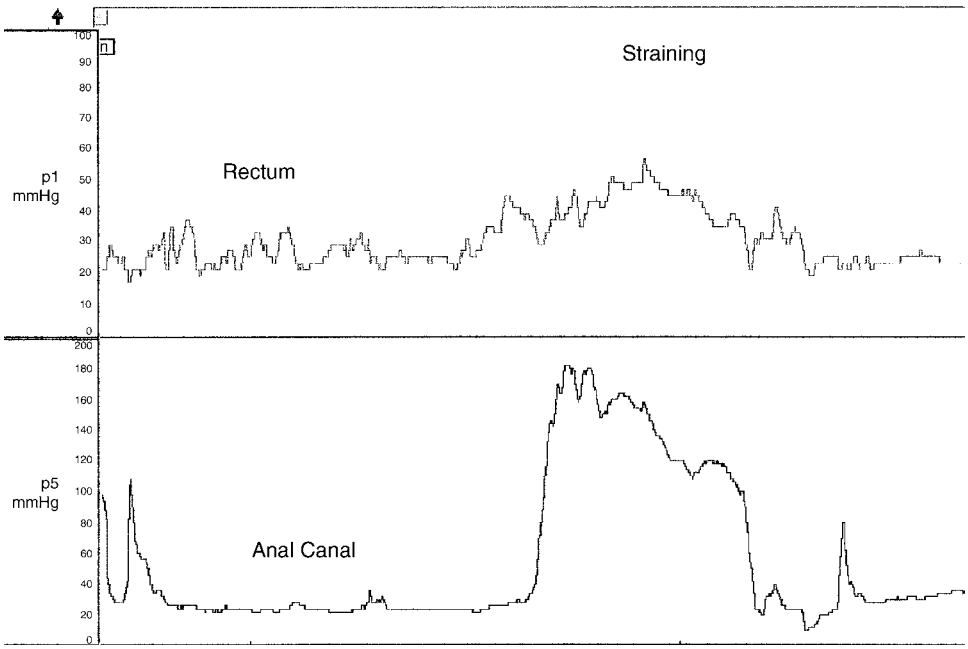


Figure 5. Dyssynergic defecation. During strain maneuver there is rise in intrarectal pressure together with a paradoxical rise in anal sphincter pressure.

addition, it provides new information that could influence the management and outcome of patients with disorder of defecation (36,37).

Anorectal manometry has gained wide acceptance as a useful method to objectively assess the physiology of defecation. However, there are some problems with anorectal physiologic testing. There is a lack of uniformity with regards to the anorectal manometry equipments, methods of performance, and interpretation of the tests. A multiplicity of catheter designs exists, including water-perfused catheters, microtransducers, and microballoons. The techniques of manometric measurement are variable. The catheter can be left at one position (stationary technique), it can be manually moved from one position to another (manual pull through technique), or it can be automatically delivered from one position to another (automatic pull-through technique). If the automated technique is selected, pressure can be recorded while the catheter is at rest or in motion. Pressure can be recorded in centimeters of H_2O, millimeters of mercury (mmHg), or kilopascals (kPa). There is also a relative lack of normative data stratified for age and gender. A more uniform method of performing these tests and interpreting the results is needed to facilitate a wider use of this technecology for the assessment of patients with anorectal disorders. Recently, experts from the American and European Motility Society have described a consensus document, where minimum standards for performing ARM have been described (13). By adopting such standards it is possible to standardize the technique globally that should help diagnosis and interpretation.

Medical Terms:

Compliance: It is defined as the capacity of the organ to stretch (expand) in response to an imposed force.

Defecation: The discharge of feces from the rectum.

Distal: Situated away from the center of the body, or from the point of origin, in contrast to proximal.

Table 3. Normal Manometric Data During Anorectal Manometry[a,b]

	All (n = 45)	Male (N = 18)	Female (N = 22)
Length of anal sphincter, cm	3.7 (3.6–3.8)	4.0 (3.8–4.2)	3.6 (3.4–3.8)
Maximum anal rest pressure, mmHg	67 (59–74)	71 (52–90)	64 (53–75)
Sustained squeeze pressure, mmHg	138 (124–152)	163 (126–200)	117 (100–134)
Squeeze duration, s	28 (25–31)	32 (26–38)	24 (20–28)
% increase in anal sphincter pressure during squeeze	126 (89–163)	158 (114–202)	103 (70–136)
Rectal pressure when squeezing, mmHg	19 (14–23)	24 (15–33)	16 (11–21)
Anal pressure during party balloon inflation, mmHg	127 (113–141)	154 (138–170)	106 (89–123)
Rectal pressure during party balloon inflation, mmHg	63 (54–72)	66 (51–81)	62 (51–73)

[a]Mean 95% cl.
[b]From Ref. 11 with permission.

Dyssynergia:	When an act is not performed smoothly or accurately because of lack of harmonious association of its various components; when there is lack of coordination or dyssynergia of the abdominal and pelvic floor muscles that are involved in defecation it is called dyssynergic defecation
ENS:	Abbreviation for enteric nervous system.
Endoanal Cushion:	Within the anus. Anal mucosal folds together with anal vascular cushion.
High pressure zone:	Intense compression area.
Intrinsic Plexus:	A network or inter-joining of nerves and blood vessels or of lymphatic vessels belonging entirely to a part.
Myenteric Plexus:	A plexus of unmyelinated fibers and postganglionic autonomic cell bodies lying in the muscular coat of the esophagus, stomach, and intestines; it communicates with the subserous and submucous plexuses, all subdivisions of the enteric plexus.
Orad:	In a direction toward the mouth.
Phasic:	In stages, in reference to rectal balloon distension for sensory testing.
Proctalgia:	Pain in the anus, or in the rectum.
Proximal:	Nearest the trunk or the point of origin, in contrast to distal.
Supraconal:	Above a condyle.
Tone:	Normal tension or resistance to stretch.

BIBLIOGRAPHY

Cited References

1. Strohbehn K. Normal pelvic floor anatomy. Obstet Gynecol Clin N Am 1998;25:683–705.
2. Whitehead WE, Schuster MM. Anorectal physiology and pathophysiology. Am J Gastroenterol 1987;82:487–497.
3. Matzel KE, Schmidt RA, Tanagho EA. Neuroanatomy of the striated muscular anal continence mechanism. Implications for the use of neurostimulation. Dis Colon Rectum 1990;33: 666–673.
4. Fernandez-Fraga X, Azpiroz F, Malagelada JR. Significance of pelvic floor muscles in anal incontinence. Gastroenterology 2002;123:1441–1450.
5. Gunterberg B, Kewenter J, Petersen I, Stener B. Anorectal function after major resections of the sacrum with bilateral or unilateral sacrifice of sacral nerves. Br J Surg 1976;63: 546–554.
6. Sun WM, Rao SS. Manometric assessment of anorectal function. Gastroenterol Clin N Am 2001;30:15–32.
7. Sun WM, Read NW. Anorectal function in normal human subjects: effect of gender. Int J Colorectal Disease 1989;4: 188–196.
8. Rao SSC. Book Chapter—Colon Transit and Anorectal Manometry. In: Rao SSC, editors. Gastrointestinal Motility: Tests and Problem-Orientated Approach. New York: Kluwer Academic/Plenum Publishers; 1999. pp 71–82.
9. Diamant NE, Kamm MA, Wald A, Whitehead WE. AGA technical review on anorectal testing techniques. Gastroenterology 1999;116:735–760.
10. McHugh SM, Diamant NE. Effect of age, gender, and parity on anal canal pressures. Contribution of impaired anal sphincter function to fecal incontinence. Dig Dis Sci 1987; 32:726–736.
11. Rao SS. Manometric tests of anorectal function in healthy adults. Am J Gastroenterol 1999;94:773–783.
12. Taylor BM, Beart RW, Jr., Phillips SF. Longitudinal and radial variations of pressure in the human anal sphincter. Gastroenterology 1984;86:693–697.
13. Rao SS. Minimum standards of anorectal manometry. Neurogastroenterol Motil 2002;14:553–559.
14. McHugh SM, Diamant NE. Anal canal pressure profile: a reappraisal as determined by rapid pullthrough technique. Gut 1987;28:1234–1241.
15. Pedersen IK, Christiansen J. A study of the physiological variation in anal manometry. Br J Surg 1989;76: 69–70.
16. Azpiroz F, Enck P, Whitehead WE. Anorectal functional testing: review of collective experience. Am J Gastroenterol 2002;97:232–240.
17. Rao SS. Dyssynergic defecation. Gastroenterol Clin N Am 2001;30:97–114.
18. MacDonagh R, et al. Anorectal function in patients with complete supraconal spinal cord lesions. Gut 1992;33:1532–1538.
19. Wald A. Colonic and anorectal motility testing in clinical practice. Am J Gastroenterol 1994;89:2109–2115.
20. Sun WM, et al. Sensory and motor responses to rectal distention vary according to rate and pattern of balloon inflation. Gastroenterology 1990;99:1008–1015.
21. Whitehead WE, Delvaux M. Standardization of barostat procedures for testing smooth muscle tone and sensory thresholds in the gastrointestinal tract. The Working Team of Glaxo-Wellcome Research, UK. Dig Dis Sci 1997;42: 223–241.
22. Mertz H, et al. Altered rectal perception is a biological marker of patients with irritable bowel syndrome. Gastroenterology 1995;109:40–52.
23. Sun WM, Read NW, Miner PB. Relation between rectal sensation and anal function in normal subjects and patients with faecal incontinence. Gut 1990;31:1056–1061.
24. Rao SS, et al. Anorectal sensitivity and responses to rectal distention in patients with ulcerative colitis. Gastroenterology 1987;93:1270–1275.
25. Salvioli B, et al. Rectal compliance, capacity, and rectoanal sensation in fecal incontinence. Am J Gastroenterol 2001;96: 2158–2168.
26. Tuteja AK, Rao SS. Review article: Recent trends in diagnosis and treatment of faecal incontinence. Aliment Pharmacol Ther 2004;19:829–840.

27. Engel AF, Kamm MA, Bartram CI, Nicholls RJ. Relationship of symptoms in faecal incontinence to specific sphincter abnormalities. Int J Colorectal Disease 1995;10:152–155.
28. Felt-Bersma RJ, Klinkenberg-Knol EC, Meuwissen SG. Anorectal function investigations in incontinent and continent patients. Differences and discriminatory value. Dis Colon Rectum 1990;33:479–485; discussion 485–486.
29. Sun WM, Donnelly TC, Read NW. Utility of a combined test of anorectal manometry, electromyography, and sensation in determining the mechanism of 'idiopathic' faecal incontinence. Gut 1992;33:807–813.
30. Sun WM, et al. Anorectal function in patients with complete spinal transection before and after sacral posterior rhizotomy. Gastroenterology 1995;108:990–998.
31. Caruana BJ, Wald A, Hinds JP, Eidelman BH. Anorectal sensory and motor function in neurogenic fecal incontinence. Comparison between multiple sclerosis and diabetes mellitus. Gastroenterology 1991;100:465–470.
32. Tjandra JJ, et al. Anorectal physiological testing in defecatory disorders: a prospective study. Aust N Z J Surg 1994;64: 322–326.
33. Rao SS, Welcher KD, Leistikow JS. Obstructive defecation: a failure of rectoanal coordination. Am J Gastroenterol 1998;93: 1042–1050.
34. Read NW, et al. Anorectal function in elderly patients with fecal impaction. Gastroenterology 1985;89:959–966.
35. Kumar D, et al. Prolonged anorectal manometry and external anal sphincter electromyography in ambulant human subjects. Dig Dis Sci 1990;35:641–648.
36. Rao SS, Patel RS. How useful are manometric tests of anorectal function in the management of defecation disorders? Am J Gastroenterol 1997;92:469–475.
37. Vaizey CJ, Kamm MA. Prospective assessment of the clinical value of anorectal investigations. Digestion 2000;61:207–214.

See also BIOFEEDBACK; ESOPHAGEAL MANOMETRY; GASTROINTESTINAL HEMORRHAGE.

ANTIBODIES, MONOCLONAL. See MONOCLONAL ANTIBODIES.

APNEA DETECTION. See VENTILATORY MONITORING.

ARRHYTHMIA, TREATMENT. See DEFIBRILLATORS; PACEMAKERS.

ARRHYTHMIA ANALYSIS, AUTOMATED

STEPHANIE A. C. SCHUCKERS
Clarkson University
Potsdam, New York

INTRODUCTION

Sudden cardiac death is estimated to affect $\sim 400{,}000$ people annually (1). Most of these cases are precipitated by ventricular fibrillation (VF), a chaotic abnormal electrical activation of the heart. Ventricular fibrillation disturbs systemic blood circulation and causes immediate death if therapy in the form of an electrical shock is not immediately applied. In fact, survival depends dramatically on the time it takes for therapy to arrive (2). Automated arrhythmia detection is a key component for speeding up defibrillation therapy through medical devices that detect arrhythmia and provide treatment automatically without human oversight. Examples of devices include implantable defibrillators, public access automated external defibrillators, and more.

Arrhythmias generally are an abnormal electrical activation of the heart. These abnormalities can occur in the atrial chambers, ventricular chambers, or both. Since the ventricles are the chambers responsible for providing blood to the body and lungs, disruptions in the electrical system that stimulates the mechanical contraction of the heart can be life threatening. Examples of ventricular arrhythmias include VF and ventricular tachycardia (VT), as seen in Fig. 1. Atrial arrhythmias including atrial fibrillation (AF), atrial flutter (AFl), and supraventricular tachycardia (SVT) are not immediately life threatening, but can cause uncomfortable symptoms and complications over the long term.

Automated arrhythmia analysis is the detection of arrhythmias through the use of a computer. This article focuses on arrhythmia detection performed *without human oversight*. The primary focus will be algorithms developed for the implantable cardioverter defibrillator (ICD). An implantable cardioverter defibrillator is a device that provides an electrical shock to ventricular fibrillation and tachycardia to terminate it and restart NSR. The implantable cardioverter defibrillator was developed in the late 1970s and FDA-approved in the mid-1980s (4–7). A

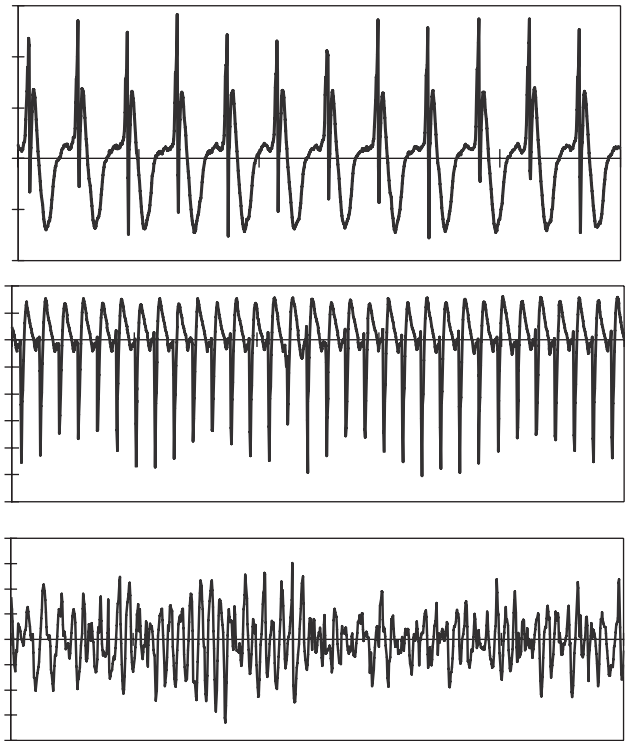


Figure 1. Unipolar electrograms (measurements of the electrical activity from inside the heart) for normal sinus rhythm (NSR), ventricular tachycardia (VT), and VF [10 s of the passage are shown (AAEL234) (3)].

Table 1. Truth Table Used to Determine Sensitivity and Specificity, Measurements of Automated Algorithm Performance

Device/Truth->	VT/VF	All Others
VT/VF	True positive	False positive
All others	False negative	True negative

catheter placed in the right ventricle is used for both sensing and therapy. This device has a long history of arrhythmia detection algorithms developed in research laboratories and brought to the marketplace. Other medical devices that use purely automated arrhythmia detection include the automatic external defibrillator and the implantable atrial defibrillator. Semiautomated arrhythmia detection is used in ambulatory and bedside monitoring. These topics will be touched on briefly.

It is important to consider the measurements used to assess the performance of automated arrhythmia analysis. Sensitivity is defined as the percent correct detection of disease, while specificity is the percent correct detection of not disease. Take the case of an implantable defibrillator that detects ventricular tachycardia and ventricular fibrillation. Consider the truth table in Table 1.

$$\text{Sensitivity} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}}$$

$$\text{Specificity} = \frac{\text{True Negatives}}{\text{True Negatives} + \text{False Positives}}$$

A false positive is one minus the specificity, while a false negative is one minus the sensitivity.

Early Work

The earliest examples of computer-based arrhythmia analysis are semiautomated approaches for bedside and ambulatory monitoring. Ambulatory monitoring typically uses a 24 or 48 h, three-lead, portable electrocardiogram (ECG) recorder that the patient wears to diagnosis arrhythmias. Arrhythmia analysis is done in an off-line fashion with technician oversight, such that it is not purely automated (8). Other ambulatory monitors include loop recorders or implanted monitors like Medtronic Reveal Insertable Loop Recorder that permanently records with patient interaction. Clinical bedside monitors typically are also not fully automated, but are used as initial alarms, which is then over read by clinical staff.

In ambulatory monitoring, in addition to detection of arrhythmias, it is typical to also detect premature ventricular contractions (PVCs). These PVCs are beats that form ectopically in the ventricle and result in an early, wide ECG beat and occur alone or in small groups of two or more. They are considered a potential sign of susceptibility to arrhythmias.

Use of correlation is a common tool in surface arrhythmia analysis (9–18). Correlation waveform analysis (CWA) uses the correlation coefficient between a previously stored template of sinus rhythm and the unknown cycle under analysis. The correlation coefficient, used by CWA, is computed as

$$\rho = \frac{\sum_{i=1}^{i=N}(t_i - \bar{t})(s_i - \bar{s})}{\sqrt{\sum_{i=1}^{i=N}(t_i - \bar{t})^2 \sum_{i=1}^{i=N}(s_i - \bar{s})^2}}$$

where ρ = the correlation coefficient, N = the number of template points, t_i = the template points, s_i = the signal points under analysis, $\bar{t}$ = the average of the template points, and $\bar{s}$ = the average of the signal points. The correlation coefficient falls within a range $-1 < \rho < +1$, where $+1$ indicates a perfectly matched signal and template.

To compute CWA, a beat detector (described in more detail in the section implantable cardioverter defibrillators) finds the location of each beat. From the location of each beat, the template is aligned with the beat under analysis, typically using the peak, and the correlation coefficient is calculated. Often, the template is shifted and the procedure is repeated to determine the best alignment indicated by the highest correlation coefficient. An example of CWA is shown in Fig. 2. Sustained high correlation indicates normal sinus rhythm and low indicates an arrhythmia.

Examples of other features used in PVC and arrhythmia detection include timing, width, height, area, offset, first spectral moment (5–25 Hz), T-wave slope, and others (9,13,14,17,20–24). Another early approach was to develop a database of electrocardiogram templates grouping similar shaped complexes based on shape, width, and prematurity (17,25–28).

Some of the earliest algorithms for *purely automated* arrhythmia detection involved algorithms for newly developing implantable devices for SVT termination and the developing implantable defibrillator (29,30). With problems of inducing ventricular arrhythmias in the devices for SVT termination, focus shifted to the implantable

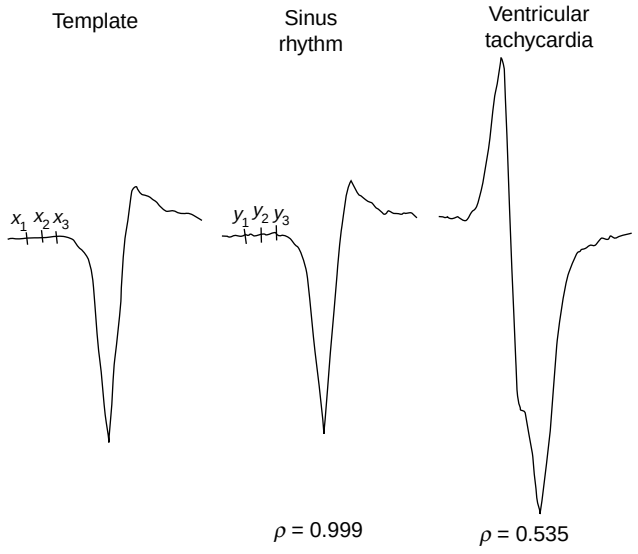


Figure 2. Example of use of correlation waveform analysis. The first electrogram is the stored normal sinus rhythm template, the second is a NSR beat with correlation equal to 0.999 and the third is a ventricular tachycardia with correlation equal to 0.535. (Used with permission from Ref. 19.)

defibrillator (31,32). In the early 1980s, Furman (29) proposed that two sensors (atrial and ventricular) be required for automatic diagnosis of tachycardia (with even a possible refinement of a third sensor for hestidine (His) bundle detection). He also suggested examining the QRS configuration for a match with sinus rhythm as a schema for diagnosing supraventricular tachycardia. Other early work included the development of algorithms for surface electrocardiography that used an esophogeal electrode for analysis of atrial information (33,34).

The original detection mechanism in the first implantable defibrillator, the AICD, was the probability density function (PDF). This algorithm utilized the derivative of the signal to define the duration of time that the signal departed from baseline (31,32) and was empirically based upon the observation that the ventricular fibrillation signal spends the majority of its time away from the electrocardiographic isoelectric baseline when compared to sinus rhythm or supraventricular rhythms (Fig. 1). The PDF was supplanted at a very early stage by intrinsic heart rate measures.

The need to identify and cardiovert ventricular tachycardia in addition to detecting and defibrillating ventricular fibrillation, and the recognition that sufficiently slow VT might have rates similar to those that may occur during sinus rhythm or supraventricular tachycardias resulted in several changes being incorporated into the second generation of devices. An alternative time-domain method called temporal electrogram analysis was incorporated into some second-generation devices (35). This algorithm employed positive and negative thresholds, or rails, placed upon electrograms sensed during sinus rhythm. A change in electrogram morphology was identified when the order of the excursion of future electrograms crossed the predetermined thresholds established during sinus rhythm. The combination of this morphologic method with ventricular rate was intended to differentiate ventricular tachycardia from other supraventricular tachycardias including sinus tachycardia.

Experience with probability density function and temporal electrogram analysis in first- and second-generation devices was disappointing. Probability density function was found to be unable to differentiate sinus tachycardia, supraventricular tachycardia, ventricular tachycardia, and ventricular fibrillation whose respective rates exceeded programmed device thresholds for tachycardia identification (36). A similar experience was encountered with temporal electrogram analysis. As a result, these criteria were utilized less and less frequently as increasing numbers of second-generation devices were implanted. By 1992, $< 15\%$ of all ICDs implanted worldwide utilized either algorithm for tachycardia discrimination (37).

IMPLANTABLE CARDIOVERTER DEFIBRILLATORS

Over time, the implantable cardioverter defibrillator added capabilities to pace terminate and cardiovert ventricular tachycardia and provide pacemaker functions, single and dual chamber. Early reviews of automated algorithms particularly for implantable defibrillators are given in Refs. 38–40. More recent reviews of automated arrhythmia detection algorithms include a thorough review by Jenkins and the author in 1996 (41) and reviews incorporating recent developments in dual chamber algorithms in 1998 and 2004 (42,43).

Rate-Based Analysis

The main method for detection of arrhythmias after initial use of PDA and TEA was the use of intrinsic heart rate for detection of ventricular tachycardia and ventricular fibrillation. To this day, all algorithms in ICDs have rate as a fundamental component for detection of arrhythmias. Since implantable defibrillators have a stable catheter screwed in the apex of right ventricle, the rate of the ventricles can be determined with little of the noise that is present at the surface of the body, like motion artifact and electromyogram noise. Ventricular tachycardia and ventricular fibrillation have rates of $\sim$ 120–240 beats per minute and $>$ 240 beats per minute, respectively.

Many approaches abound for arrhythmia detection using rate, but the general procedure is the same (Fig. 3). First, each ECG beat must be detected. Second, the time between beats (or the cycle length) is determined. Most algorithms rely on cycle length (CL) values over beats per minute. The value of the CL determines the zone it falls into: Normal, VT, or VF. In some cases, zones may be further divided depending on the device and therapy options. The thresholds that define the zones are programmable. Each zone has a programmable counter that will determine when therapy will need to be considered. In addition, each zone has a reset mechanism that may be different depending on the zone. For example, typical VT zones require X consecutive beats within the zone or the counter is reset. While VF often has an X of Y criteria, for example, 12 of 16 beats. This flexibility is due to the fact that VF is of varying amplitude, morphology, and rate, such that each beat may not be detected reliably and/or may not be in the VF zone. The CL thresholds, counters, and associated therapies are all programmable.

The fundamental basis of automated rate algorithms is the detection of each beat. Many approaches have been suggested and utilized including fixed thresholds, exponentially varying thresholds, amplitude gain control, and

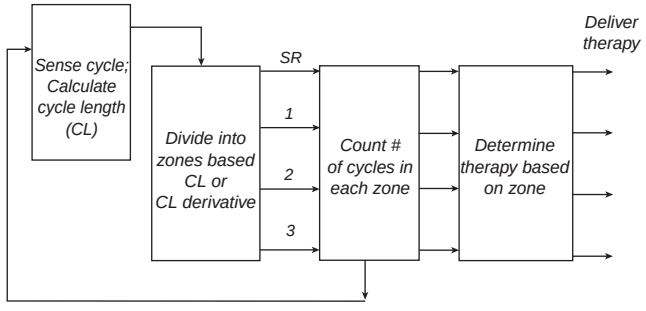


Figure 3. Typical rate-based arrhythmia detection scheme for implantable cardioverter defibrillators. First, each beat is detected and the cycle length between beats determined. A counter is incremented in the zone that the CL falls and therapy is delivered when the counter reaches a programmed threshold.

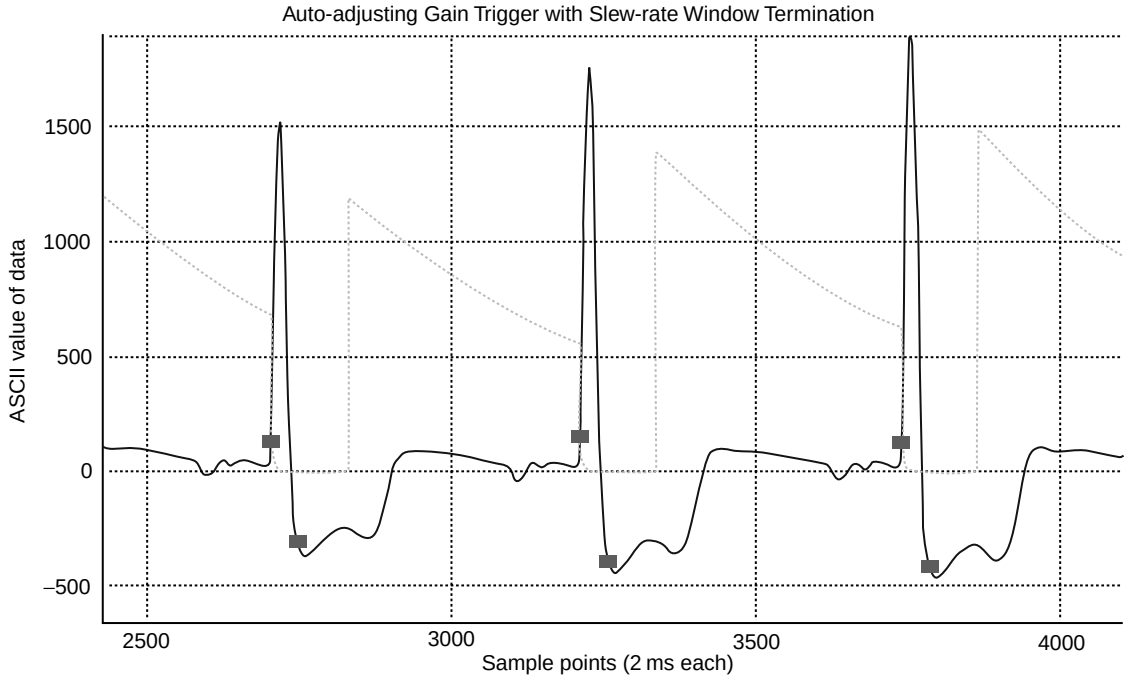


Figure 4. Example of beat detector that utilizes an exponentially decaying threshold. After a beat is detected, a blanking period prevents detection of the same beat twice. Then, the threshold (dotted line) for determination of the next beat is calculated as a percentage of the peak amplitude of the previous beat. This threshold exponentially decays, such that beats that are smaller than the currently detected beat will be not be missed (47).

others (9,44–46). In implantable devices, most are hardware-based and beyond the scope of this article. For example, one software method relies on an exponentially varying threshold (Fig. 4) (47). After a beat is detected, there is first a blanking period to prevent a beat from being detected more than once. After the blanking period, the threshold for detection of the next beat is set as a percentage of the previous beat. This threshold then decays exponentially such that subsequent beats that have a smaller peak amplitude will be detected. Most beat detectors also have a floor for the threshold, that is, the smallest amplitude by which a beat can be detected to prevent detection of noise as a beat.

An example of one rate-based algorithm is given in Fig. 5 (40). This algorithm uses three CL thresholds, fibrillation detection interval (FDI), fast tachycardia interval (FTI), and tachycardia detection interval (TDI), and two counters, VF counter (VFCNT) and VT counter (VTCNT). These are combined to result in three zones, VF, fast VT, and slow VT, which can have different therapeutic settings, utilizing defibrillation shock, cardioversion, and antitachycardia pacing. Therapy for ventricular fibrillation is given when 18 of 24 beats are shorter than the FDI. Therapy for slow ventricular tachycardia is delivered when 16 beats counted by the VTCNT are between the FDI and TDI thresholds. The VTCNT will be reset by one long CL greater than TDI. The fast VT zone is a combination of these techniques.

A thorough description of the rate-based algorithms is given in Ref. 40. While many additional features have been added to refine the decision, the main structure of automated arrhythmia detection algorithms still rely on this fundamental approach (42).

As can be see from Fig. 1, heart rate in VT and VF increase substantially over normal sinus rhythm. This is a reliable means of detecting VT and VF for implantable devices, resulting in high sensitivity. Unfortunately, while providing high sensitivity, heart rate also increases for normal reasons, exercise, stress, resulting in sinus tachycardia or for nonventricular-based arrhythmias like atrial fibrillation, supraventricular tachycardia, atrial flutter, and so on, which do not require therapy from the ICD. Thus, rate-based algorithms have low specificity. False therapies have been estimated in as much as 10–40% in the early devices (48–50). Morphology and other extended algorithmic approaches have long been suggested as a means to increase specificity.

Early rate-based algorithms to prevent false therapies, due to sinus tachycardia, atrial fibrillation, and so on, include onset and rate stability. Rate-based methods were chosen initially over morphology due to the simplicity of calculations in battery operated devices.

Onset is the difference between the rate changes during the onset of sinus tachycardia compared to those of VT, since the onset of VT is typically sudden compared to sinus tachycardia. False therapies due to sinus tachycardia are determined by onset. Figure 6a shows the sudden onset of ventricular tachycardia.

Rate stability is used to prevent false therapies due to AF. In AF, it is common for the ventricle to respond to the atrium at a fast rate. This response is typically irregular since atrial fibrillation, by definition has an irregular rate,

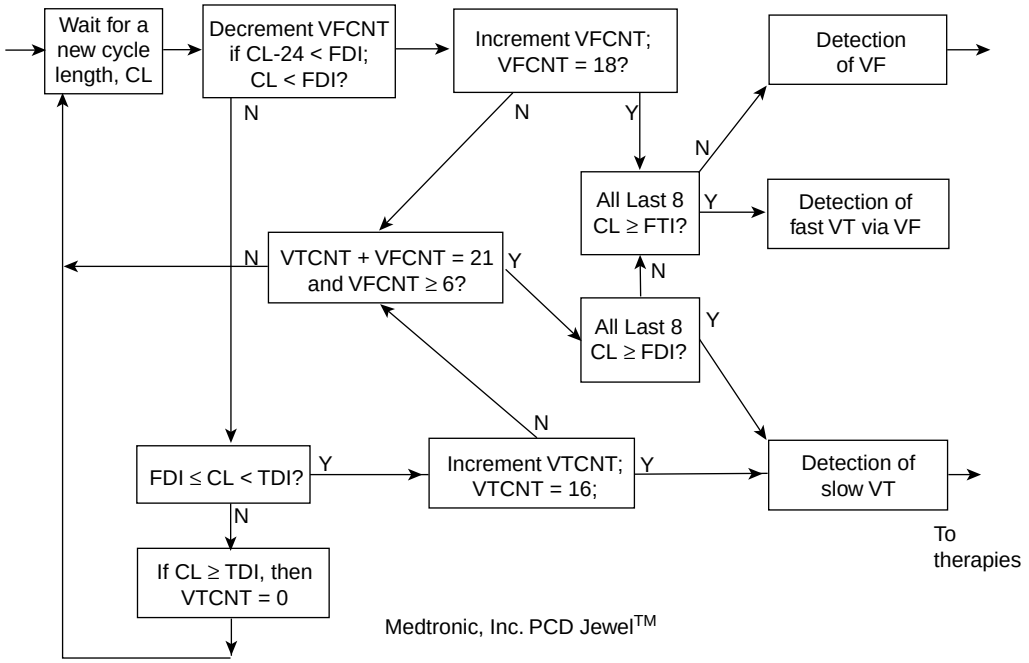


Figure 5. Example of rate-based algorithm for the Medtronic PCD Jewel. The VFCNT is the VF counter, FDI is the fibrillation detection interval, FTI is the fast tachycardia interval, VTCNT is the VT counter, and TDI is the tachycardia detection interval. This algorithm has three zones that has associated programmable therapies including defibrillation shock, cardioversion, and antitachycardia pacing (40).

and since not every beat is conducted from the atrium to the ventricle. Rate stability considers the stability of the ventricular rate, since VT typically has a stable rate compared to the ventricular response to atrial fibrillation. Figure 6b shows an irregular ventricular response to atrial flutter.

Rate and rate-derived measures that measure onset and stability (based on cycle-by-cycle interval measurements) include average or median cycle length, rapid deviation in cycle length (onset), minimal deviation of cycle length (stability), and relative timing measures in one or both chambers or from multiple electrodes within one or more chambers. Among the methods most widely used for detection of VT in commercially available single chamber antitachycardia devices have been combinations of rate, rate stability, and sudden onset (51–56). Pless and Sweeney published an algorithm for (1) sudden onset, (2) rate stability, and (3) sustained high rate (57). This schema among others (58,59) was a forerunner of many of the methods introduced into tachycardia detection by ICDs (60).

Morphological Pattern Recognition

Instead of relying purely on rate, it has been suggested that morphology may provide the means for automated arrhythmia detection to separate VT and VF from rhythms with fast rates that do not need therapy. Morphology in this context refers to characteristics of the electrogram waveform itself, which are easily identifiable and measurable. Such features might include peak-to-peak amplitude, slew rate (a measure of waveform), sequence of slope patterns, sequence of amplitude threshold crossings, statistical pattern recognition of total waveform shape by correlation coefficient measures, and others (61,62). Figure 1 shows an example of distinctly different waveforms recorded from the right ventricular apex during SR, VT, and VF (3). Furthermore, morphology in the ventricle appears normal

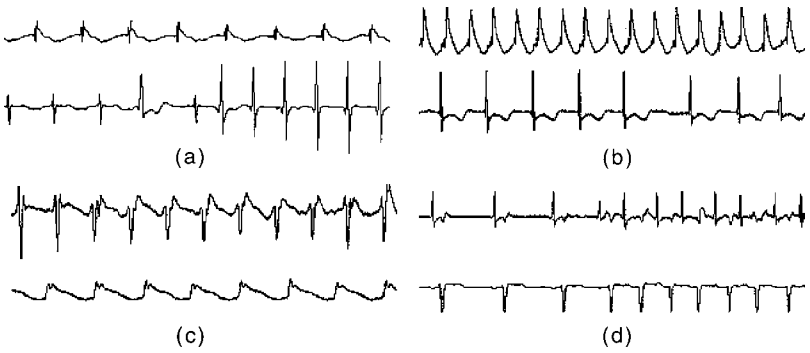


Figure 6. Atrial (top) and ventricular (bottom) electrograms. (a) Sudden onset of VT with normal atrial electrogram, (b) irregular response of ventricular to atrial flutter, (c) simultaneous atrial flutter and ventricular tachycardia, and (d) sudden onset of supraventricular tachycardia with ventricular response.

even during supraventricular arrhythmias since the rhythm typically is conducted normally in the ventricles.

The ICDs often have two channels of electrograms. The first channel is typically bipolar, which is associated with two electrodes on the lead, or an electrode–coil combination, both located within the ventricle. This channel provides a near-field electrical view of the ventricle and is usually used for beat detection because the electrogram typically has a narrow QRS (the main depolarization of the electrogram). The second electrode configuration is far-field which, for example, may use an electrode in the ventricle versus the implantable device casing. The far-field electrode combination is used primarily for giving the electrical shock. However, this far-field view typically has a more global perspective of the electrogram depolarization and is helpful in differentiating the morphology changes between normal beats and ventricular abnormalities.

Template-Based Algorithms

Correlation Waveform Analysis. Lin et al. (19,62,63) investigated three techniques for morphologic analysis of VT: correlation waveform analysis, amplitude distribution analysis, and spectral analysis. Correlation waveform analysis (CWA) is a classic method of pattern recognition applied to the surface electrocardiogram, described earlier, but was first applied to intracardiac signals in this study. Correlation waveform analysis was shown to be superior and has the advantage of being independent of amplitude and baseline fluctuations. However, it requires heavy computational demands. Less computationally demanding algorithms based on the same principle have been developed and are described in the next section.

Less Computationally Demanding Template-Based Algorithms. Another template matching algorithm based on raw signal analysis measured the area of difference between electrograms, that is, adding absolute values of the algebraic differences between each point on the electrogram and corresponding point on the SR template (64,65). The area of difference was expressed as a percentage of the total area of the template. The measurement of an area of difference is simple computationally, but has the disadvantage of producing erroneous results in the face of baseline and amplitude fluctuations, and this method fails to produce a bounded measure. An improvement on this technique by signal normalization and scaling to create a metric bounded by ± 1 was utilized by Throne et al. (66).

Steinhaus et al. (67) modified correlation analysis of electrograms to address computational demand by applying data compression to filtered data (1–11 Hz) by retaining only samples with maximum excursion from the last saved sample. The average squared correlation coefficient (ρ^2) was used for separation of SR and VT. Comparison with noncompressed correlations demonstrated that data compression had negligible effects on the results.

Throne et al. (66) designed four fast algorithms and compared discrimination results to CWA performance. These morphological methods were the bin area method (BAM); derivative area method (DAM); accumulated difference of slopes (ADIOS); and normalized area of difference (NAD). All four techniques are independent of amplitude fluctuations and three of the four are independent of baseline changes.

The bin area method is a template matching algorithm that compares corresponding area segments or bins of the template with the signal to be analyzed. Each bin (average of three consecutive points) is adjusted for baseline fluctuations by subtracting the average of the bins over one cycle and normalized to eliminate amplitude variations. This BAM equation is given in the following equation:

$$\rho = 1 - \sum_{i=1}^{i=M} \left| \frac{T_i - \overline{T}}{\sum_{k=1}^{k=M} \left| T_k - \overline{T} \right|} - \frac{S_i - \overline{S}}{\sum_{k=1}^{k=M} \left| S_k - \overline{S} \right|} \right|$$

where the bins are $S_1 = s_1 + s_2 + s_3$, $S_2 = s_4 + s_5 + s_6, \cdots$, $S_M = s_{N-2} + s_{N-1} + s_N$ and the average of M bins is $\overline{S} = (1/M)\Sigma_{k=1}^{k=M} S_k$. The bins and average of the bins is calculated similarly for the template. The BAM metric falls between -1 and $+1$, allowing a comparison to CWA.

Normalized area of difference is identical to BAM except that the average bin value is not removed. By not removing the average value the algorithm avoids one division that would otherwise increase computational demand each time the BAM algorithm is applied. The NAD is independent of amplitude changes.

The DAM uses the first derivative of the template and the signal under analysis. The method creates segments from zero crossings of the derivative of the template. It imposes the same segmentation for analysis of the derivative of the signal to be compared. The segments are normalized, but are not adjusted for baseline variations since derivatives are by their nature baseline independent. The DAM metric is calculated as follows:

$$\rho = 1 - \sum_{i=1}^{i=M} \left| \frac{\dot{T}_i}{\sum_{k=1}^{k=M} \left| \dot{T}_k \right|} - \frac{\dot{S}_i}{\sum_{k=1}^{k=M} \left| \dot{S}_k \right|} \right|$$

where $\dot{T}_k$ represents the kth bin of the first derivative of the template. The DAM metric falls between -1 and $+1$.

The ADIOS is similar to DAM in that it also employs the first derivative of the waveforms. A template is constructed of the sign of the derivative of the ventricular depolarization template. This template of signs is then compared to the signs of the derivative for subsequent depolarizations. The total number of sign differences between the template and the current ventricular depolarization is then computed as

$$\rho = \sum_{i=1}^{i=N} \text{sign}(\dot{t}_i) \oplus \text{sign}(\dot{s}_i)$$

where $\oplus$ is the exclusive or operator. The number of sign changes is bounded by 0 and the maximum number of points in the template (N), that is, $\rho \in \{0, \ldots, N\}$.

Evaluation of these four algorithms was performed on 19 patients with 31 distinct ventricular tachycardia morphologies. Three of the algorithms (BAM, DAM, and

NAD) performed as well or better than correlation waveform analysis, but with one-half to one-tenth the computational demands.

A morphological scheme for analysis of ventricular electrograms (SIG) was devised for minimal computation (68) and compared to NAD. The SIG is a template-based method that creates a boundary window enclosing all template points that form a signature of the waveform to be compared. Equivalent results of VT separation were seen in the two techniques at two thresholds, but at an increased safety margin of separation SIG outperformed NAD and yielded a fourfold reduction in computation.

Another simplified correlation-type algorithm has been designed using electrogram vector timing and correlation, developed for the Guidant ICD (69). In this algorithm, the rate (near-field) channel is used for determining the location of each beat. The peak of the near-field electrogram or fiducial point is used for alignment of the template with the beat under analysis. From this fiducial point, eight specific points are chosen on the shock (far-field) electrogram. The amplitude of the shock channel at the rate-channel fiducial point is one point. In addition, amplitudes at the turning point, intermediate, and baseline values on the shock channel are selected as shown in Fig. 7. This provides an eight-point template that is compared to subsequent beats using the square of the correlation coefficient, as follows:

$$\mathrm{FCC} = \frac{(8\sum t_i s_i - (\sum t_i)(\sum s_i))^2}{(8\sum t_i^2 - (\sum t_i)^2)(8\sum s_i^2 - (\sum s_i)^2)}$$

where each summation is $i = 1$–8.

When an unknown beat is analyzed, the exact same timing relative to the fiducial point as the template is used for selecting the amplitudes of the unknown beat. For beats that have a different morphology, those points will not be associated with the same amplitudes as the normal template beat and, thus, the correlation coefficient will be low. To incorporate this into an overall scheme to detect an arrhythmia, morphology was calculated for a sliding window of 10 beats. If 8 or more beats were detected as abnormal, a VT was detected.

Another algorithm that reduces computational complexity of the standard correlation algorithm uses the wavelet transform of the sinus beat for the template (70). Wavelets can reduce the number of coefficients needed to characterize a beat while still retaining the important morphologic information. The sinus electrogram is transformed using the Haar (square) wavelet, considering a family of 48 wavelets over 187.5 ms window aligned by the fiducial point of the QRS. The wavelet transform is simplified by removing the standard factor of square root of 2. In addition, wavelet coefficients that do not carry much information, defined by a threshold, are set to zero. The remaining coefficients are normalized. This gives a variable template size, depending on the electrogram, but typically between 8 and 20 coefficients. To analyze an unknown electrogram, the electrogram is aligned using the peak (negative or positive) point. The wavelet transform is computed for the unknown electrogram and each coefficient is compared using the absolute difference in wavelet coefficients (c_i) between the template and unknown beat. A match is determined by the following equation:

$$\mathrm{Match}\% = \left|1 - \frac{\sum |c_i^{\mathrm{template}} - c_i^{\mathrm{unknown}}|}{\sum |c_i^{\mathrm{template}}|}\right| * 100$$

The nominal threshold used in this study is 70%. This morphology algorithm is incorporated into an overall rate scheme by remaining inactive until a ventricular tachycardia has been detected by the rate algorithm. Then, the morphology is calculated for the preceding eight beats. A VT is detected if six or more beats are detected as abnormal.

A novel way of testing this algorithm was used. Instead of, as in most tests, using data prerecorded in laboratory conditions, this algorithm was downloaded to the Medtronic clinical ICDs and tested off-line, while the device functioned with its regular algorithm.

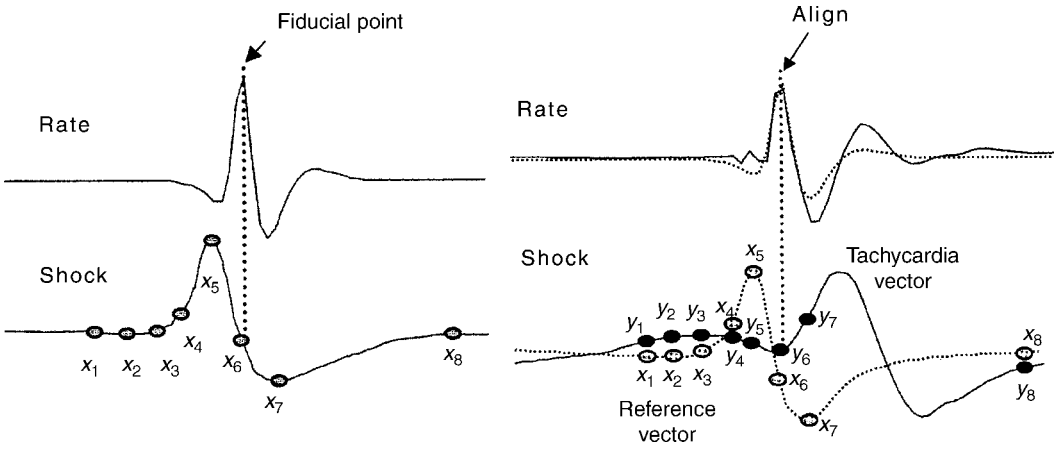


Figure 7. Example for vector timing and correlation algorithm. Alignment of the template is based on the peak of the rate electrogram channel. From the peak, eight specific points on the shock channel are automatically selected for the template (left). These exact points in time relative to the fiducial point selected from the template are applied to the "unknown" beat (right). (Used with permission from Ref. 69.)

Another algorithm that utilizes morphology is termed morphology discrimination (MD) in St. Jude implantable cardioverter defibrillators (71,72). This algorithm uses an area-based approach. First, the template is defined based on the three consecutive peaks with the largest area. The area under each peak is normalized by the maximum peak area. When analyzing an unknown beat, the beat is aligned with the template using the dominant peak of the unknown beat. If this peak does not have same polarity, the second largest peak is used. If this also does not have the same polarity, a nonmatch is declared. Once the unknown beat and template are aligned, the morphology score is determined by the following:

$$\text{Score} = (1 - ||N\text{Area}A - N\text{Area}A'| + |N\text{Area}B - N\text{Area}B'| + |N\text{Area}C - N\text{Area Peak}C'||) * 100$$

where NArea stands for the normalized area of the three corresponding peaks of the template (A, B, C) and test complexes (A', B', C'). For arrhythmia diagnosis, once the rate criteria is met, the algorithm determines the number of matching complexes in the morphology window. If the number of matching complexes equals or exceeds a programmed number of matching complexes, VT is not confirmed and therapy is not delivered. This is repeated for as long as the rate criteria has been met or VT is confirmed.

Template matching by CWA was further examined for distinction of multiple VTs of unique morphologies in the same patient (73,74). It was hypothesized that, in addition to a SR template, a second template acquired from the clinical VT could provide confirmation of a later recurrence of the same VT. The recognition of two or more different VTs within the same patient could play an important role in future devices in the selection of therapy to be delivered to hemodynamically stable versus unstable VTs.

Considerations for Template Analysis. While template-based algorithms appear the most promising, several issues need to be addressed. The first is that it is necessary that the normal sinus rhythm beat or template remain stable, that is, does not change over time or due to position or activity. Several studies using temporary electrodes saw changes in the morphology of normal rhythm due to increase in rate or positional changes (75–78) but further studies with fixed electrodes showed no changes in morphology due to heart rate or position, with some changes in amplitude (76,79). A second consideration is that paroxysmal (sudden) bundle branch block (BBB) may be misdiagnosed as ventricular tachycardia (80). While this may result in a false therapy, it does not result in withholding of therapy during life-threatening arrhythmias (the more critical mistake).

Feature-Based Algorithms

Depolarization Width for Detection of Ventricular Tachycardia. Depolarization width (i.e., duration) in ventricular electrograms has been used as a discriminant of supraventricular rhythm (SR) from VT (81,82). Electrogram width is available in the Medtronic single chamber ICDs. This criterion uses a slew threshold to find the beginning and end of the QRS. Analysis of electrogram width compared to a patient-specific width threshold is performed using the previous eight beats after a VT detected by the rate component of the algorithm. If a minimum of six complexes are greater than the width, then a VT is detected. Otherwise, the counter is reset. This algorithm is not appropriate in patients with BBB that have a wider width for normal beats. Exercise induced variation should be considered in programming (83). Electrogram width has been shown to be sensitive to body position and changes over longer periods of time (6 months in this study) (84).

Amplitude and Frequency Analysis. Amplitude and frequency are distinguishing characteristics of arrhythmia. Amplitude during ventricular tachycardia is typically higher and during ventricular fibrillation is lower than normal sinus rhythm (85,86). These differences have not been considered pronounced and consistent enough, such that a classifier could be based on them.

Frequency-domain analysis is often proposed for classification of rhythms (87) but little success has been solidly demonstrated for the recognition of VT (63). Distinctly different morphological waveforms (SR vs VT), which are easily classified in the time domain, can exhibit similar or identical frequency components if one focuses on the depolarization component alone. Examination of longer segments of 1000–15,000 ms yields the same phenomenon because the power present in small visually distinctive high frequency notches is insignificant compared to the remainder of the signal, and changes in polarity of the waveform, easily recognized in the time domain, are simply not revealed by frequency analysis (63). Frequency-domain recognition of AF (88) and VF (89,90) is perhaps more promising. However, frequency has not been applied in commercial applications given the success of rate and time-domain morphology approaches.

Other Morphologic Approaches. Other approaches that have been suggested in the literature include use of neural networks (91–97). Neural network approaches utilize either features, the time-series, or frequency components as inputs to the neural network. The network is trained on one dataset and tested on a second. Limitations with the approaches developed thus far are related to the fact that there is only limited data for development of the neural network. One problem is that in some studies the training set and test set both include samples from the same patient. Thus, these networks cannot be considered a general classifier for all patients, since it did not have a valid test set for assessing results on unseen patients. Ideally, three sets should be utilized: training, validation, and testing. The purpose of the validation set is to test the generalization of the network, such that it is not overtrained. Plus, it is typical practice to retrain neural networks until good results are achieved on the validation set. A separate testing set verifies that success on the validation set was not just by chance. Until large datasets are available for development of the algorithms, neural networks will not be considered for clinical use. Furthermore, neural networks generally have not achieved much

acceptance by the clinical community who prefer methods that are tied to underlying physiologic understanding.

Dual-Chamber Arrhythmia Detection

Since dual-chamber pacemakers have been combined into ICDs, the possibility of the use of information from the atrial electrogram for arrhythmia diagnosis has opened up. The most prevalent cause of delivery of false therapy is AF, which accounts for > 60% of all false shocks according to the literature. The simple addition of an atrial sensing lead can dramatically change the false detection statistics.

The first two-channel algorithm for intracardiac analysis incorporated timing of atrial activation as well as ventricular into the diagnostic logic of arrhythmia classification (98,99). This scheme was based on earlier work in which an esophageal pill electrode (33) provided P-wave identification as an adjunct to surface leads in coronary care and Holter monitoring (34). The early argument for adding atrial sensing for improvement of ICD tachycardia detection was advanced conceptually by Furman in 1982 (29), was demonstrated algorithmically by Arzbaecher et al. in 1984 (58), and was further confirmed by Schuger (100). This simple two-channel analysis offers a first-pass method for confirming a VT diagnosis when the ventricular rate exceeds the atrial (Fig. 8). Recognition of a run of short intervals was followed by a comparison of atrial versus ventricular rate. With both chambers (atrial and ventricular) under analysis, most supraventricular arrhythmias could be detected by an $N : 1$ ($A : V$) relationship, and most ventricular arrhythmias could be detected by a $1 : N$ ($A : V$) relationship. Ambiguity occurred in tachycardias characterized by a 1 : 1 relationship, where SVT with 1 : 1 ventricular conduction could be confounded with ventricular tachycardia with retrograde 1 : 1 atrial conduction. In addition, an $N : 1$ ($A : V$) relationship should not be an automatic detection of atrial arrhythmia, since a concurrent ventricular arrhythmia could be masked by a faster atrial arrhythmia, as seen in Fig. 6c. Thus the limitations of two-channel timing analysis, although powerful, needs to be addressed by more advanced logical relationships.

A system designed for two-channel analysis using rate in both chambers plus three supplemental time features (onset derived by median filtering, regularity, and multiplicity) was designed for real-time diagnosis (101) of spontaneous rhythms. This system was an integration of previously tested stand-alone timing schemes (102,103). The combined system is able to recognize competing atrial and ventricular tachycardias and produces joint diagnoses of the concurrent rhythms. Simultaneous VT and atrial flutter is classified via atrial rate, ventricular rate, and a lack of multiplicity. Fast ventricular response in AF is detected via the regularity criterion. Onset (employed in 1:1 tachycardias) utilizes a median filter technique (102).

Commercially, each manufacturer now has available algorithms that utilize information from both chambers for making the diagnosis, particularly for improving specificity. These algorithms are implemented in commercial devices and continually updated and improved. Examples of the algorithms are in the following paragraphs. Reviews are given in Refs. 42,43 along with a thorough comparison of clinical results of the various commercial dual-chamber algorithms (43). Other comparisons include Refs. 104,105.

The first actual realization of a two-channel ICD appeared with the introduction into clinical trials (1995) of the ELA *Defender,* a dual chamber sensing and pacing ICD that uses both atrial and ventricular signals for its tachycardia diagnoses (106) (Fig. 9). The first step after a fast rate is detected is to consider stability of the ventricular rate. If the rhythm is not stable, atrial fibrillation is detected and no therapy delivered. The next consideration is the association between the A and V. For $A : V$ association of $1 : N$ or no association, a VT is detected. For $N : 1$ association, atrial arrhythmia is detected and no therapy delivered. For 1:1 association, the last step is consideration of chamber of onset, ventricular acceleration will result in VT therapy, while no acceleration or atrial acceleration will result in no therapy. An example of a sudden onset in the atrium due to SVT is seen in Fig. 6d. The most recent algorithm, PARAD+ incorporates additional features after the association criteria (second step) (107). If there is no PR association, a second criteria is considered where a single

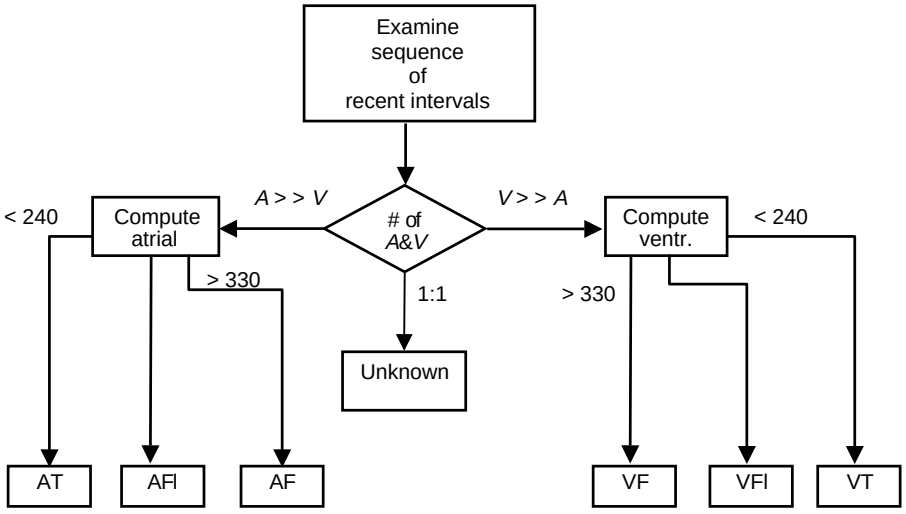


Figure 8. Basic dual chamber arrhythmia detection algorithm. For a sequence of intervals, the number of atrial (A) intervals is compared to the number of ventricular (V) intervals. If there are more V than A, a diagnosis of ventricular fibrillation (VF), ventricular flutter (VFl), or ventricular tachycardia (VT) is made based on the rate. If there are more V than A beats, a diagnosis of atrial fibrillation (AF), atrial flutter (AFl), or atrial tachycardia is made (AT) (58).

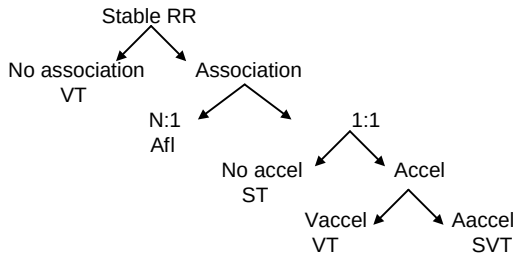


Figure 9. Flowchart of dual-chamber arrhythmia detection algorithm using simple rate-based features. For unstable RR interval (time between beats), atrial fibrillation is detected. For stable RR and no association between the atrium (A)and ventricle (V), ventricular tachycardia (VT) is detected. For $N:1$ ($A:V$) association, atrial flutter (AFl) is detected. For $1:1$ ($A:V$) association with no acceleration (Accel), sinus tachycardia (ST) is detected. Last, with a ventricular acceleration, VT is detected and with atrial acceleration (Aaccel) supraventricular tachycardia (SVT) is detected (106).

long ventricular cycle will result in the diagnosis of atrial fibrillation (for the next 24 consecutive cycles) while no long ventricular cycles will result in VT detection.

The Guidant Ventak AV III DR algorithm uses the following scheme, shown in Fig. 10 (104). First, it checks if the ventricular rate is greater than the atrial rate (by 10 bpm). If yes, then VT is detected. If no, then more analysis is performed. If the atrial rate is greater than the atrial fibrillation threshold and the RR intervals are not stable, then supraventricular rhythm is classified. If the RR intervals are stable, VT is detected. If the atrial rate is not greater than the atrial fibrillation threshold, then ventricular tachycardia is detected if the RR intervals are stable and there is a sudden onset of ventricular rate. An updated algorithm from Guidant is described in the next section.

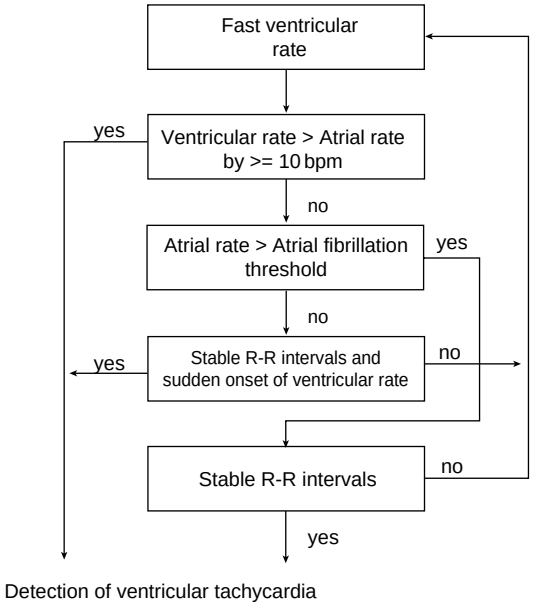


Figure 10. Algorithm for Guidant Ventak AV III DR, using comparison of the rate of A and V, stability and onset. (Used with permission from Ref. 104.)

Perhaps the most complex of the dual chamber algorithms rests with the PR Logic algorithm, by Medtronic (42). This algorithm uses a series of measurements from the timing of the atrial ventricular depolarizations to create a code (1 of 19 possible). These codes are then used for ultimate diagnosis. The main component of the algorithm is the timing between the atrial and ventricular beat to determine if the conduction was antegrade or retrograde. For a given ventricular RR interval, if the atrial beat falls 80 ms before or 50 ms after the ventricular beat, the rhythm is considered junctional. Outside of this, if the atrial beat (P-wave) falls within the first one-half of the RR interval, it is considered retrograde conduction. If the atrial beat falls in the second half of the RR interval it is considered antegrade. This is performed for the previous two beats and incorporated in the code. There are only a few programmable components in the algorithm. The first is the type of SVT for which rejection rules apply (AF/AFL, ST, SVT). The second is the SVT limit. The rest are not programmable.

Dual-Chamber with Ventricular Morphological Analysis

The Photon DR from St. Jude incorporates morphology in its dual chamber defibrillator algorithm. The MD in the ventricular chamber described earlier is incorporated in the full algorithm as follows (108). For $V > A$, ventricular tachycardia is detected. For $V < A$, a combination of morphology discrimination and interval stability is used to inhibit therapy for atrial fibrillation/flutter and SVT. For the branch $V = A$, morphology discrimination and sudden onset is used to inhibit therapy for ST and SVT. This algorithm has an automatic template feature update (ATU) for real-time calibration of the sinus template.

A new dual chamber algorithm from Guidant uses the vector timing and correlation (VTC) algorithm, described earlier (69). If the V rate exceeds A rate by $>$ 10 bpm, a VT is detected. Otherwise, VTC algorithm is implemented. If the atrial rate does not exceed the AF threshold, then VTC will be used for diagnosis. Otherwise, stability will be used for diagnosis. Therapy would be inhibited for an unstable ventricular rhythm.

Two-Channel Morphological Analysis. An early algorithm that uses morphological analysis of both the intraatrial signal and the intraventricular signal (109,110) is based on strategy developed previously for surface and esophageal signals (111). A five-feature vector was derived for each cycle containing an atrial and a ventricular waveform metric (ρ_a, ρ_v), where ρ is the correlation coefficient for each depolarization, and AA, AV, and VV interval classifiers (short, normal, and long). Single-cycle codes were mapped to 122 diagnostic statements. The eight most current cycles were employed for a contextual interpretation of the underlying rhythm. This addition of morphological analysis of both atrial and ventricular channels combined with rate determination in each channel on a cycle-by-cycle basis, dramatically demonstrated the power of modern signal processing in the interpretation of arrhythmias.

One aspect in which analysis of the atrial morphology would be very useful in ICDs is the separation of antegrade versus retrograde atrial conduction. During a 1:1 tachycardia, it is difficult to separate an SVT with 1:1

anterograde conduction (forward conduction from the sinus node through the atrium and AV node to the ventricle) versus a ventricular arrhythmia with retrograde conduction (retrograde conduction from the ventricle through the AV node to the atrium). To differentiate these cases, morphology differences in the atrial electrogram could be utilized, where abnormal morphology would indicate retrograde conduction. Various methods have been described in the literature which use similar approaches as ventricular morphology (112–118).

Distinction of Ventricular Tachycardia and Ventricular Fibrillation

Discriminating between VT and VF might be useful to allow unique zone settings for choice of therapy. Antitachycardia pacing is a lower energy therapy used to treat VT, which is not painful to the patient. Currently, there is difficulty in detecting each VF cycle, leading to electrogram dropout, which leads physicians to expand the VF detection zone to eliminate the possibility of misdiagnosing VF (119,120). Therefore, many VTs are detected as VF and given shock therapy directly. While these are typically fast VTs, there is a possibility that fast VTs can be terminated using anti-tachycardia pacing protocols, with only limited delay of shock therapy, if fast VTs and VF could be differentiated (121). In one study, 76% of fast VTs would have received shock therapy if programmed traditionally (121). However, by expanding the fast VT zone, 81% diagnosed as fast VT were effectively pace-terminated. More sophisticated digital signal processing techniques could be applied to ensure proper separation of VT and VF by methods more intelligent than counting alone.

For separation of VT and VF, CWA using a sinus rhythm template was tested on a passage of monomorphic ventricular tachycardia and a subsequent passage of ventricular fibrillation in each patient (122–124). The standard deviation of the correlation coefficient (ρ) of each class (VT and VF) was used as a discriminant function. This scheme was based upon the empiric knowledge that correlation values are more tightly clustered in the cycle-by-cycle analysis of monomorphic VT and more broadly distributed in the dissimilar waveforms in VF. Results showed easy separation of sinus rhythm from VT and VF; however in the VT/VF separation, standard deviation only achieved limited success Standard deviation requires patient-specific thresholds, may not hold for all template-based algorithms, and adds further computational requirements to the algorithm; therefore, it is not a promising algorithm in its present form for discrimination of VT from VF.

Throne et al. (125) addressed the problem of separating monomorphic and polymorphic VT/VF by using scatter diagram analysis. A moving average filter was applied to rate and morphology channels and plotted as corresponding pairs of points on a scatter diagram with a 15×15 grid. The percentage of grid blocks occupied by at least one sample point was determined. Investigators found that monomorphic VTs trace nearly the same path in two-dimensional space and occupy a smaller percentage of the graph than nonregular rhythms such as polymorphic VT or VF.

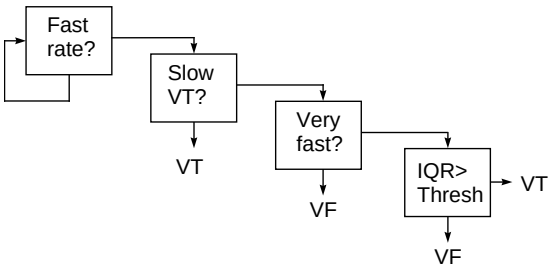


Figure 11. Basic algorithm for separation of VT and VF using PSC. Once a fast rate is detected, VT and VF are detected for slow fast rates and very fast rates, respectively. Only in the overlap between fast VT and VF rates, is the morphology algorithm implemented. Interquartile range (IQR) of the paired signal concordance over a passage is used.

A magnitude-squared coherence function was developed by Ropella et al. (126), which utilizes filtering and Fourier transformation of intraventricular electrograms derived from two leads with a sliding window to distinguish monomorphic ventricular tachycardia from polymorphic ventricular tachycardia and ventricular fibrillation. This method, while elegant, requires multiple electrode sites and is at present too computationally demanding for consideration in battery operated devices. As technology advances, the possibility of hardware implementation of frequency-based methods such as magnitude-squared coherence and time-domain CWA may become feasible.

A similar algorithm uses two "unipolar" ventricular electrograms, 1 cm apart, to compare the paired signal concordance (PSC) between the electrograms using correlation analysis (127). During normal rhythms and VT, the two closely spaced electrograms will exhibit high correlation, while during VT, the two electrogram will experience low correlation. Considering only rhythms that have a fast rate in the overlap between fast VT and VF rates, the variability of the correlation, measured by interquartile range, over a passage distinguishes VT from VF (Fig. 11).

Complexity measurements have also been utilized for distinction of ventricular tachycardia and fibrillation, including approximate entropy (128), Lempil–Ziv complexity (129), least-squares prony modeling algorithm (130).

OTHER DEVICES THAT USE AUTOMATED ARRHYTHMIA DETECTION

Other commercially available devices that use automated arrhythmia detection algorithms include the automatic external defibrillator and the implantable atrial defibrillator.

Automatic External Defibrillators

Recently, automatic external defibrillators (AEDs) have become widespread and available. The AED is able to determine if the rhythm for an unresponsive, pulseless patients is shockable or unshockable and is able to apply therapy automatically or to inform the user to deliver the

therapy (131–133). The AEDs are available on location for large organizations, such as airports, airplanes, businesses, sporting events, schools, and malls (134). This expanded availability dramatically increases the possibility that victims of ventricular fibrillation could receive defibrillation in a timely manner, thus, improving survival rates (135).

The AEDs, operating in a truly automated mode, must be exquisitely accurate in their interpretation of the ECG signal (136,137). In an AED, shockable rhythms are rhythms that will result in death if not treated immediately and include coarse ventricular fibrillation and ventricular tachycardia. Nonshockable rhythms are rhythms where no benefit and even possible harm may result from therapy and include supraventricular tachycardia, atrial fibrillation normal sinus rhythm, and asystole. Asystole is not considered shockable for these devices since the leads may be misplaced and no signal captured. Intermediate rhythms are rhythms that may or may not receive benefit from therapy and include fine VF and VT. Ventricular tachycardia is an intermediate rhythm because often it is hemodynamically tolerated in the patient. A rate threshold is usually programmed in the device (even though there is still no universally accepted or obvious delineation in rate for hemodynamic tolerance in the literature).

Contrary to the ICD, AEDs have an extremely necessary requirement for accurate specificity (shocks when not needed) since these devices are expected to be used by untrained personnel. Algorithms must consider large variations in cardiac rhythms and artifact from CPR or patient movement. The risk to the patient and the technician is too great to allow public use for devices with decreased specificity. Given that a bystander is on the scene and that trained help may soon be available, specificity is more important for the first, or immediate response. However, sensitivity must be considered to ensure the AEDs potential to save lives is maximized. In the AED, more battery power can be utilized (since the device is not implanted), and therefore more sophisticated schemes borrowed from ICD technology have been considered.

Current devices use numerous schemes for determining if the patient is in ventricular fibrillation. Common components include isoelectric content (like PDF algorithm of the ICD), zero crossings, rate, variability, slope, amplitude and frequency (138), all similar techniques to those described in the ICD literature. A review of AED algorithms is given in Ref. 138.

One example of a recent algorithm described in the literature is for a programmable automatic external defibrillator designed to be used in the hospital setting for monitoring and automatic defibrillation, if needed (139). This device uses a programmable rate criterion to detect shockable rhythms. In addition, the device has an algorithm which distinguishes VT and SVT rhythms below a SVT threshold. The algorithm uses three features to discriminate between SVT and VT. The first is the modulation domain function that uses amplitude and frequency characteristics. The second, called waveform factor (WF), provides a running average of the electrocardiogram signal amplitude normalized by the R-wave amplitude. The WF for one beat is as follows:

$$\mathrm{WF}_i = \frac{100 \times \sum_{1}^{N} \mathrm{abs}(A_n)}{N \times \mathrm{abs}(A_r)}$$

where N is the total number of samples between the previous and current beat, A_n is the nth amplitude of the signal, and A_r is the peak. The algorithm uses an eight beat running average. The SVT rhythms would have a small WF value, while VT (which has a wide QRS complex on the surface of the body) would have a high WF value. This algorithm should not be used with patients who have bundle branch block, chronic or paroxysmal. The third feature is called amplitude variability analysis factor, which uses distribution of the average derivative. The measurement is found by calculating the number of derivatives which fall into the baseline bin as a percentage of the total number of sample points. Amplitude variability analysis (AVA) is calculated as follows:

$$\mathrm{AVA} = \frac{100 \times \sum n(i)}{N}$$

where the summation is performed across the baseline bins and $n(i)$ is the number of samples for the ith bin. The baseline bins are defined as 25% of the total bins centered at $n(i)_{\max}$. Exact use of these features in the AED algorithm are not described.

Implantable Atrial Defibrillators

Implantable atrial defibrillators are used in patients with paroxysmal or persistent atrial fibrillation, particularly those which are symptomatic and drug refractory (140,141). Goals in atrial defibrillators are different than ICDs since atrial tachyarrhythmias are typically hemodynamically tolerable, therefore, more time and care can be used to make the decision. The challenge is that the device must sense low, variable amplitude atrial signals, while not sensing far-field ventricular waves. Furthermore, there are also multiple therapies available, antitachycardia pacing for atrial tachycardia, and cardioversion for atrial fibrillation. Lastly, some atrial defibrillators have been combined with ICDs such that back-up ventricular defibrillation therapy is available in this susceptible population (140,142).

A review of algorithms used in atrial defibrillators in given in Ref. 143. For example, Medtronic has a dual-chamber defibrillator that has both atrial and ventricular therapies (140,142). The algorithm for detection uses the same algorithm as the dual-chamber ventricular (only) defibrillator, PR Logic. In addition to this algorithm, there are two zones used for detection of atrial tachycardia and of atrial fibrillation. If the zones overlap, AT is detected if it is regular and AF if it is irregular. The purpose of multiple zones is similar to ventricular devices, in that a variety of therapies can be selected and utilized for each zone. For this device, this includes pacing algorithms for prevention, pacing therapies for termination, and high voltage shocks.

CONCLUSION

This article focuses on the overall approaches used for automated arrhythmia detection. However, this review did not delve into the specifics of comparisons of sensitivity and specificity results for the various algorithms. While, each paper referenced gives performance for a specific test database, it is difficult to compare the results from one study to another. There have been some attempts to develop standardized datasets, including surface electrocardiograms from Physionet (including the MIT-BIH databases) (144) and American Heart Association (145), and intracardiac electrograms, in addition to surface, from Ann Arbor Electrogram Libraries (3). Use of these datasets allows for comparisons, but does not address the differences between performance at the system level that incorporates the hardware components of the specific device. A comprehensive description of the pitfalls in comparing results from one study to another is given in (43). These include limitations of (*1*) benchtesting that does not incorporate specific ICD-system differences and spontaneous arrhythmias, (*2*) limited storage in the ICD making gold standard clinical diagnosis difficult, (*3*) great variations in settings of rate based thresholds and zones, (*4*) variability of types of rhythms included in the study, among others.

In conclusion, examples from the long history of automated arrhythmia detection for implantable cardioverter defibrillators is given with a brief mention of automated external defibrillators and implantable atrial defibrillators. The ICDs are beginning to reach maturity in terms of addressing both sensitivity and specificity in performance of the algorithms to achieve close to perfect detection of life-threatening arrhythmias, with greatly reduced false therapies. In the meantime, automated external defibrillators and implantable atrial defibrillators have learned many lessons from the ICD experience to provide accurate arrhythmia diagnosis. Devices on the horizon incorporating automated arrhythmia detection may include wearable external defibrillators (146,147), wearable wireless monitors, and beyond. This rich area of devices that detect and treat life-threatening arrhythmias shall reduce the risk of sudden cardiac death.

BIBLIOGRAPHY

Cited References

1. American Heart Association. Heart Disease and Stroke Statistics—2005 Update. American Heart Association. Available at http://www.americanheart.org. Accessed 2005.
2. Vilke GM, et al. The three-phase model of cardiac arrest as applied to ventricular fibrillation in a large, urban emergency medical services system. Resuscitation 2005;64:341–346.
3. Jenkins JM, Jenkins RE. Arrhythmia database for algorithm testing: surface leads plus intracardiac leads for validation. J Electrocardiol 2003;36(Suppl):157–1610. Available at http://www.electrogram.com.
4. Nisam S, Barold S. Historical evolution of the automatic implantable cardioverter defibrillator in the treatment of malignant ventricular tachyarrhythmias. In: Alt E, Klein H, Griffin JC, editors. The implantable cardioverter/defibrillator. Berlin: Springer-Verlag; 1992. pt. 1, p 3–23.
5. Mower MM, Reid PR. Historical development of automatic implantable cardioverter-defibrillator. In: Naccarelli GV, Veltri EP, editors. Implantable Cardioverter-Defibrillators, Boston: Scientific Publications; 1993. Chapt. 2. p 15–25.
6. Mower MJ. Clinical and historical perspective. In: Singer I, editor. Implantable Cardioverter Defibrillator. Armonk (NY): Futura; 1994, Chapt. 1. p 3–12.
7. Bach SM, Shapland JE. Engineering aspects of implantable defibrillators. In: Saksena S, Goldschlager N, editors. Electrical Therapy for Cardiac Arrhythmias. Philadelphia: Saunders; 1990. Chapt. 18. p 371–383.
8. Kennedy HL. Ambulatory (Holter) Electrocardiography Technology. Cardiol Clin 1992;10:341–359.
9. Feldman CL, Amazeen PG, Klein MD, Lown B. Computer detection of ventricular ectopic beats. Comput Biomed Res 1970 Dec; 3(6):666–674.
10. Thomas LJ, et al. Automated Cardiac Dysrhythmia Analysis. Proc IEEE Sept 1979;67(9):1322–1337.
11. Collins SM, Arzbaecher RC. An efficient algorithm for waveform analysis using the correlation coefficient. Comput Biomed Res 1981 Aug; 14(4):381–389.
12. Hulting J, Nygards ME. Evaluation of a computer-based system for detecting ventricular arrhythmias. Acta Med Scand 1976;199(12):53–60.
13. Thakor NV. From Holter monitors to automatic defibrillators: developments in ambulatory arrhythmia monitoring. IEEE Trans Biomed Eng 1984 Dec; 31(12):770–778.
14. Feldman CL, Hubelbank M. Cardiovascular Monitoring in the Coronary Care Unit. Med Instrum 1977;11:288–292.
15. Hubelbank M, Feldman CL. A 60x computer-based Holter tape processing system. Med Instrum 1978;Nov–Dec; 12(6):324–326.
16. Govrin O, Sadeh D, Akselrod S, Abboud S. Cross-correlation technique for arrhythmia detection using PR and PP intervals. Comput Biomed Res 1985 Feb; 18(1):37–45.
17. Shah PM, et al. Automatic real time arrhythmia monitoring in the intensive coronary care unit. Am J Cardiol 1977 May 4; 39(5):701–708.
18. Lipschultz A. Computerized arrhythmia monitoring systems: a review. J Clin Eng 1982 Jul–Sep; 7(3):229–234.
19. Lin D, et al. Analysis of time and frequency domain patterns of endocardial electrograms to distinguish ventricular tachycardia from sinus rhythm. Comp Cardiol 1987; 171–174.
20. Knoebel SB, Lovelace DE, Rasmussen S, Wash SE. Computer detection of premature ventricular complexes: a modified approach. Am J Cardiol 1976 Oct; 38(4):440–447.
21. Yanowitz F, Kinias P, Rawling D, Fozzard HA. Accuracy of a continuous real-time ECG dysrhythmia monitoring system. Circulation. 1974 July; 50(1):65–72.
22. Mead CN, et al. A detection algorithm for multiform premature ventricular contractions. Med Instrum 1978;12:337–339.
23. Knoebel SB, Lovelace DE. A two-dimensional clustering technique for identification of multiform ventricular complexes. Med Instrum 1978;12:332–333.
24. Cheng QL, Lee HS, Thakor NV. ECG waveform analysis by significant point extraction. II. Pattern matching. Comput Biomed Res 1987 Oct; 20(5):428–442.
25. Spitz AL, Harrison DC. Automated family classification in ambulatory arrhythmia monitoring. Med Instrum 1978 Nov–Dec; 12(6):322–323.
26. Oliver GC, et al. Detection of premature ventricular contractions with a clinical system for monitoring electrocardiographic rhythms. Comput Biomed Res 1971 Oct; 4(5): 523–541.
27. Yanowitz F, Kinias P, Rawling D, Fozzard HA. Accuracy of a continuous real-time ECG dysrhythmia monitoring system. Circulation 1974 July; 50(1):65–72.

28. Cooper DH, Kennedy HL, Lyyski DS, Sprague MK. Holter triage ambulatory ECG analysis. Accuracy and time efficiency. J Electrocardiol. 1996 Jan; 29(1):33–38.
29. Furman S, Fisher JK, Panizzo F. Necessity of signal processing in tachycardia detection. In: Barold SS, Mugica J, editors. The Third Decade of Cardiac Pacing: Advances in Technology and Clinical Applications. Mt Kisco (NY): Futura; 1982. Pt. 3, Chapt. 1. p 265–274.
30. Jenkins J, et al. Present state of industrial development of devices. PACE May–June 1984;7(II):557–568.
31. Mirowski M, Mower MM, Reid PR. The automatic implantable defibrillator. Am Heart J 1980;100:1089–1092.
32. Langer A, Heilman MS, Mower MM, Mirowski M. Considerations in the development of the automatic implantable defibrillator. Med Instrum May–June 1976;10:163–167.
33. Arzbaecher R. A pill electrode for the study of cardiac dysrhythmia. Med Instrum 1978;12:277–281.
34. Jenkins JM, Wu D, Arzbaecher R. Computer diagnosis of supraventricular and ventricular arrhythmias. Circulation 1979;60:977–987.
35. Paul VE, O'Nunain S, Malik M. Temporal electrogram analysis: algorithm development. PACE Dec. 1990;13:1943–1947.
36. Toivonen L, Viitasalo M, Jarvinen A. The performance of the probability density function in differentiating supraventricular from ventricular rhythms. PACE May 1992;15: 726–730.
37. DiCarlo L, et al. Tachycardia detection by antitachycardia devices: present limitations and future strategies. J Interven Cardiol 1994;7:459–472.
38. Pannizzo F, Mercando AD, Fisher JD, Furman S. Automatic methods for detection of tachyarrhythmias by antitachycardia devices. PACE Feb. 1988;11:308–316.
39. Lang DJ, Bach SM. Algorithms for fibrillation and tachyarrhythmia detection. J Electrocardiol 1990;23(Suppl):46–50.
40. Olson WH. Tachyarrhythmia sensing and detection. In: Singer I, editor. Implantable Cardioverter Defibrillator. Armonk (NY): Futura; 1994. Chapt. 4. p 71–107.
41. Jenkins JM, Caswell SA. Detection algorithms in implantable defibrillators. Proc IEEE 1996;84:428–445.
42. Olson WH. Dual chamber sensing and detection for implantable cardioverter-defibrillators. In: Singer I, Barold SS, Camm AJ, editors. Nonpharmacological Therapy of Arrhythmias for the 21st century. Armonk (NY): Futura; 1998. p 385–421.
43. Aliot E, Mitzsche R, Ribart A. Arrhythmia detection by dual-chamber implantable cardioverter defibrillators: a review of current algorithms. Europace 2004;6:273–286.
44. Thakor NV, Webster JG. Design and evaluation of QRS and noise detectors for ambulatory ECG monitors. Med Biol Eng Comput 1982;20:709–714.
45. Jalaleddine S, Hutchens C. Ambulatory ECG wave detection for automated analysis: a review. ISA Trans 1987;26(4): 33–43.
46. Warren JA, et al. Implantable cardioverter defibrillators. Proc IEEE 1996;84:468–479.
47. MacDonald R, Jenkins J, Arzbaecher R, Throne R. A software trigger for intracardiac waveform detection with automatic threshold adjustment. Proc Computers Cardiol IEEE 30276-6574 1990; 167–170.
48. Winkle RA, et al. Long-term outcome with the automatic implantable cardioverter-defibrillator. J Am Coll Cardiol May 1989;13:1353–1361.
49. Grimm W, Flores BF, Marchlinski FE. Electrocardiographically documented unnecessary, spontaneous shocks in 241 patients with implantable cardioverter defibrillators. PACE Nov. 1992;15:670–669.
50. Nunain SO, et al. Limitations and late complications of third-generation automatic cardioverter-defibrillators. Circulation April 15 1995;91:2204–2213.
51. Warren J, Martin RO. Clinical evaluation of automatic tachycardia diagnosis by an implanted device. PACE 1986; 9:16.
52. Nathan AW, Creamer JE, Davies DW. Clinical experience with a new versatile, software based, tachycardia reversion pacemaker. J Am Coll Cardiol 1987;7:184A.
53. Olson W, Bardy G, Mehra R. Onset and stability for ventricular tachycardia detection in an implantable pacer-cardioverter-defibrillator. Comp Cardiol 1987;34:167–170.
54. Tomaselli G, Scheinman M, Griffin J. The utility of timing algorithms for distinguishing ventricular from supraventricular tachycardias. PACE March–April 1987;10:415.
55. Geibel A, Zehender M, Brugada P. Changes in cycle length at the onset of sustained tachycardias-importance for antitachycardia pacing. Am Heart J March 1988;115:588–592.
56. Ripley KL, Bump TE, Arzbaecher RC. Evaluation of techniques for recognition of ventricular arrhythmias by implanted devices. IEEE Trans Biomed Eng June 1989;36:618–624.
57. Pless BD, Sweeney MB. Discrimination of supraventricular tachycardia from sinus tachycardia of overlapping cycle length. PACE Nov–Dec 1984;7:1318–1324.
58. Arzbaecher R, et al. Automatic tachycardia recognition. PACE May–June 1984;7:541–547.
59. Jenkins JM, et al. Tachycardia detection in implantable antitachycardia devices. PACE Nov–Dec 1984;7:1273–1277.
60. Swerdlow CD, et al. Discrimination of ventricular tachycardia from sinus tachycardia and atrial fibrillation in a tiered-therapy cardioverter. J Am Coll Cardiol 1994;23:1342–1355.
61. Pannizzo F, Furman S. Pattern recognition for tachycardia detection: a comparison of methods. PACE July 1987;10:999.
62. Santel D, Mehra R, Olson W. Integrative algorithm for detection of ventricular tachyarrhythmias from the intracardiac electrogram. Comp Cardiol 1987; 175–177.
63. Lin D, DiCarlo LA, Jenkins JM. Identification of ventricular tachycardia using intracavity ventricular electrograms: analysis of time and frequency domain patterns. PACE Nov. 1988;1592–1606.
64. Tomaselli GF, et al. Morphologic differences of the endocardial electrogram in beats of sinus and ventricular origin. PACE Mar. 1988;11:254–262.
65. Langberg JL, Gibb WJ, Auslander DM, Griffin JC. Identification of ventricular tachycardia with use of the morphology of the endocardial electrogram. Circulation June 1988;77: 1363–1369.
66. Throne RD, Jenkins JM, Winston SA, DiCarlo LA. A comparison of four new time domain methods for discriminating monomorphic ventricular tachycardia from sinus rhythm using ventricular waveform morphology. IEEE Trans Biomed Eng June 1991;38:561–570. (U. S. Pat. No. 5,000,189 Mar. 19, 1991).
67. Steinhaus BM, et al. Detection of ventricular tachycardia using scanning correlation analysis. PACE Dec. 1990;13: 1930–1936.
68. Greenhut SE, et al. Separation of ventricular tachycardia from sinus rhythm using a practical, real-time template matching computer system. PACE Nov. 1992;15:2146–2153.
69. Gold MR, et al. Advanced rhythm discrimination for implantable cardioverter defibrillators using electrogram vector timing and correlation. J Cardiovasc Electrophysiol 2002; 13:1092–1097.
70. Swerdlow CD, et al. Discrimination of ventricular tachycardia from supraventricular tachycardia by a downloaded wavelet transform morphology algorithm. J Cardiovasc Electrophysiol 2002;13:432–441.

71. Duru F, et al. Morphology discriminator feature for enhanced ventricular tachycardia discrimination in implantable cardioverter defibrillators. PACE 2000;23:1365–1374.
72. Boriani G, et al. Clinical evaluation of morphology discrimination: an algorithm for rhythm discrimination in cardioverter defibrillators. PACE 2001;24:994–1001.
73. Throne RD, Jenkins JM, Winston SA, DiCarlo LA. Use of tachycardia templates for recognition of recurrent monomorphic ventricular tachycardia. Comp Cardiol 1989;171–174.
74. Stevenson SA, Jenkins JM, DiCarlo LA. Analysis of the intraventricular electrogram for differentiation of distinct monomorphic ventricular arrhythmias. J Am Coll Cardiol (submitted June 1995;).
75. Paul VE, et al. Variability of the intracardiac electrogram: effect on specificity of tachycardia detection. PACE Dec. 1990;13:1925–1829.
76. Finelli CJ, et al. Intraventraventricular electrogram morphology: effect of increased heart rate with and without accompanying changes in sympathetic tone. Comp Cardiol 1990; 115–118.
77. Rosenheck S, Schmaltz S, Kadish AH, Morady F. Effect of rate augmentation and isoproterenol on the amplitude of atrial and ventricular electrograms. Am J Cardiol July 1 1990;66:101–102.
78. Belz MK, et al. The effect of left ventricular intracavitary volume on the unipolar electrogram. PACE Sept. 1993;16: 1842–1852.
79. Caswell SA, et al. Chronic bipolar electrograms are stable during changes in body position and activity: implications for antitachycardia devices. PACE April 1995;18:871.
80. Throne RD, Jenkins JM, Winston SA, DiCarlo LA. Paroxysmal bundle branch block of supraventricular origin: a possible source of misdiagnosis in detecting ventricular tachycardia using ventricular electrogram morphology. PACE April 1990;13:453–458.
81. Gilberg JM, Olson WH, Bardy GH, Mader SJ. Electrogram width algorithms for discrimination of supraventricular rhythm from ventricular tachycardia. PACE April 1994;17:866.
82. Unterberg C, et al. Long-term clinical experience with the EGM width detection criteria for differentiation of supraventricular and ventricular tachycardia in patients with implantable cardioverter defibrillators. PACE 2000;23: 1611–1617.
83. Kingenheben T, Sticherling C, Skupin M, Hohnloser SH. Intracardiac QRS electrogram width—an arrhythmia detection feature for implantable cardioverter defibrillators: exercise induced variation as a base for device programming. PACE 1998;21:1609–1617.
84. Favale S, et al. Electrogram width parameter analysis in implantable cardioverter defibrillators: influence of body position and electrode configuration. PACE 2001;24:1732–1738.
85. Leitch JW, et al. Correlation between the ventricular electrogram amplitude in sinus rhythm and in ventricular fibrillation. PACE Sept. 1990;13:1105–1109.
86. Ellenbogen KA, et al. Measurement of ventricular electrogram amplitude during intraoperative induction of ventricular tachyarrhythmias. Am J Cardiol Oct. 15 1992;70:1017–1022.
87. Pannizzo F, Furman S. Frequency spectra of ventricular tachycardia and sinus rhythm in human intracardiac electrograms: application to tachycardia detection for cardiac pacemakers. IEEE Trans Biomed Eng June 1988; 421–425.
88. Slocum J, Sahakian A, Swiryn S. Computer discrimination of atrial fibrillation and regular atrial rhythms from intra-atrial electrograms. PACE May 1988;11:610–621.
89. Aubert AE, et al. Frequency analysis of VF episodes during AICD implantation. PACE June 1988;11(Suppl):891.
90. Lovett EG, Ropella KM. Autoregressive spread-spectral analysis of intracardiac electrograms: comparison of Fourier analysis. Comp Cardiol 1992; 503–506.
91. Minami K, Nakajima H, Toyoshima T. Real-time discrimination of ventricular tachyarrhythmia with Fourier-transform neural network. IEEE Trans Biomed Eng 1999;46:179–185.
92. Yan MC, Jenkins JM, DiCarlo LA. Feasibility of arrhythmia recognition by antitachycardia devices using time-frequency analysis with neural network classification. PACE 1995;18: 871.
93. Farrugia S, Yee H, Nickolls P. Implantable cardioverter defibrillator electrogram recognition with a multilayer perceptron. PACE Jan. 1993;16:228–234.
94. Leong PH, Jabri MA. MATIC-An intracardiac tachycardia classification system. PACE Sept. 1982;15:1317–1331.
95. Rojo-Alvarez JL, et al. Automatic discrimination between supraventricular and ventricular tachycardia using a multilayer perceptron in implantable cardioverter defibrillators. PACE 2002;25:1599–1604.
96. Wang Y, et al. A short-time multifractal approach for arrhythmia detection based on fuzzy neural network. IEEE Trans Biomed Eng 2001;48:989–995.
97. Al-Fahoum AS, Howitt I. Combined wavelet transformation and radial basis neural networks for classifying life-threatening cardiac arrhythmias. Med Biol Eng Comp 1999;27: 566–573.
98. Arzbaecher R, et al. Automatic tachycardia recognition. PACE May–June 1984;7:541–547.
99. Jenkins JM, et al. Tachycardia detection in implantable antitachycardia devices. PACE Nov–Dec 1984;7:1273–1277.
100. Schuger CD, Jackson K, Steinman RT, Lehmann MH. Atrial sensing to augment ventricular tachycardia detection by the automatic implantable cardioverter defibrillator: a utility study. PACE Oct. 1988;11:1456–1463.
101. Caswell SA, DiCarlo LA, Chiang CJ, Jenkins JM. Automated analysis of spontaneously occurring arrhythmias by implantable devices: limitation of using rate and timing features alone. J Electrocardiol 1994;27(Suppl):151–156.
102. Chiang CJ, Jenkins JM, DiCarlo LA. Discrimination of ventricular tachycardia from sinus tachycardia by antitachycardia devices: value of median filtering. Med Engr Phys Nov. 1994;16:513–517.
103. Chiang CJ, Jenkins JM, DiCarlo LA. The value of rate regularity and multiplicity measures to detect ventricular tachycardia in atrial fibrillation of flutter with a fast ventricular response. PACE Sept. 1994;17:1503–1508.
104. Hintringer F, Schwarzacher S, Eibl G, Pachinger O. Inappropriate detection of supraventricular arrhythmias by implantable dual chamber defibrillators: a comparison of four different algorithms. PACE 2001;24:835–841.
105. Hintringer F, et al. Comparison of the specificity of implantable dual chamber defibrillator detection algorithms. PACE 2004;27:976–982.
106. Lavergne T, et al. Preliminary clinical experience with the first dual chamber pacemaker defibrillator. PACE 1997;20: 182–188.
107. Mletzko R, et al. Enhanced specificity of a dual chamber ICD arrhythmia detection algorithm by rate stability criteria. PACE 2004;27:1113–1119.
108. Bailin SJ, et al. Clinical investigation of a new dual-chamber implantable cardioverter defibrillator with improved rhythm discrimination capabilities. J Cardiovasc Electrophysiol 2003;14:144–149.
109. Chiang CJ, et al. Real-time arrhythmia identification from automated analysis of intraatrial and intraventricular electrograms. PACE Jan. 1993;16:223–227.

110. Caswell SA, et al. Pattern recognition of cardiac arrhythmias using two intracardiac channels. Comp Cardiol 1993; 181–184.
111. DiCarlo LA, Lin D, Jenkins JM. Automated interpretation of cardiac arrhythmias. J Electrocardiol Jan. 1993;26:53–67.
112. Amikan S, Furman S. A comparison of antegrade and retrograde atrial depolarization in the electrogram. PACE May 1983;6:A111.
113. Wainwright R, Davies W, Tooley M. Ideal atrial lead positioning to detect retrograde atrial depolarization by digitization and slope analysis of the atrial electrogram. PACE Nov–Dec. 1984;7:1152–1157.
114. Davies DW, Wainwright RJ, Tooley MA. Detection of pathological tachycardia by analysis of electrogram morphology. PACE March–April 1986;9:200–208.
115. McAlister HF, et al. Atrial electrogram analysis: antegrade versus retrograde. PACE Nov. 1988;11:1703–1707.
116. Throne RD, et al. Discrimination of retrograde from anterograde atrial activation using intracardiac electrogram waveform analysis. PACE Oct. 1989;12:1622–1630.
117. Saba S, et al. Use of correlation waveform analysis in discrmination between anterograde and retrograde atrial electrograms during ventricular tachycardia. J Cardiovasc Electrophysiol 2001;12:145–149.
118. Strauss D, Jung J, Rieder A, Manoli Y. Classification of endocardial electrograms using adapted wavelet packets and neural networks. Ann Biomed Eng 2001;29:483–492.
119. DiCarlo LA, et al. Impact of varying electrogram amplitude sensing threshold upon the performance of rate algorithms for ventricular fibrillation detection. Circulation Oct. 1994;90:I–176.
120. Caswell SA, et al. Ventricular tachycardia versus ventricular fibrillation: Discrimination by antitachycardia devices. J Electrocardiol 1996;28:29.
121. Wathen MS, et al. PainFREE Rx II Investigators. Prospective randomized multicenter trial of empirical antitachycardia pacing versus shocks for spontaneous rapid ventricular tachycardia in patients with implantable cardioverter-defibrillators: Pacing Fast Ventricular Tachycardia Reduces Shock Therapies (PainFREE Rx II) trial results. Circulation 2004;110:2591–2596.
122. Jenkins JM, Kriegler C, DiCarlo LA. Discrimination of ventricular tachycardia from ventricular fibrillation using intracardiac electrogram analysis. PACE April 1991;14:718.
123. DiCarlo LA, Jenkins JM, Winston SA, Kriegler C. Differentiaion of ventricular tachycardia from ventricular fibrillation using intraventricular electrogram morphology. Am J Cardiol Sept. 15 1992;70:820–822.
124. Jenkins JM, Caswell SA, Yan MC, DiCarlo LA. Is waveform analysis a viable consideration for implantable devices given its computational demand? Comp Cardiol 1993; 839–842.
125. Throne RD, et al. Scatter diagram analysis: a new technique for discriminating ventricular tachyarrhythmias. PACE July 1994;17:1267–1275.
126. Ropella KM, Baerman JM, Sahakian AV, Swiryn S. Differentiation of ventricular tachyarrhythmias. Circulation Dec. 1990;82:2035–2043.
127. Caswell SA, Jenkins JM, DiCarlo LA. Comprehensive scheme for detection of ventricular fibrillation for implantable cardioverter defibrillators. J Electrocardiol 1998;30: 131–136.
128. Schuckers SA. Use of approximate entropy measurements to classify ventricular tachycardia and fibrillation. J Electrocardiol 1998;31(Suppl):101–105.
129. Zhang HX, Zhu YX, Wang ZM. Complexity measure and complexity rate information based detection of ventricular tachycardia and fibrillation. Med Biol Eng Comp 2000;38: 553–557.
130. Chen SW. A two-stage discrmination of cardiac arrhythmias using a total least squares-based prony modeling algorithm. IEEE Trans Biomed Eng 2000;47:1317–1327.
131. American Heart Assocation. Emergency Cardiac Care Committee and Subcommittees. Guidelines for cardiopulmonary resuscitation and emergency cardiac care. JAMA 1992;268: 2171–2302.
132. Aronson AL, Haggar B. The automatic external defibrillator-pacemaker: clinical rationale and engineering design. Med Instrum 1986;20:27–35.
133. Charbonnier FM. External defibrillators and emergency external pacemekers. Proc IEEE 1996;84:487–499.
134. Weisfeldt ML, et al. American Heart Association Report on the Public Access Defibrillation Conference December 8–10, 1994. Automatic External Defibrillation Task Force. Circulation 1995;92:2740–2747.
135. Dimmit MA, Griffiths SE. What's new in prehospital care? Nursing 1992;22:58–61.
136. Association for the Advancement of Medical Instrumentation. Automatic external defibrillators and remote-control defibrillators [American National Standard]. AAMI 1993; ANSI/AAMI DF39-1993.
137. American Heart Association. AED Task Force, Subcommittee on Safety and Efficacy. Automatic External Defibrillators for Public Access Use: Recommendations for Specifying and Reporting Arrhythmia Analysis Algorithm Performance, Incorporating new Waveforms, and Enhancing Safety. AHA 1996.
138. Charbonnier FM. Algorithms for arrhythmia analysis in AEDs. In: Tacker WA Jr, editor. Defibrillation of the Heart: ICDs, AEDs and Manual. St Louis (MO): Mosby/Yearbook; 1994.
139. Mattioni T, et al. Performance of an automatic external cardioverter-defibrillator algorithm in discrimination of supraventricular from ventricular tachycardia. Am J Cardiol 2003;91:1323–1326.
140. Sopher SM, Camm AJ. Atrial defibrillators. In: Singer I, Barold SS, Camm AJ, editors. Nonpharmacological Therapy of Arrhythmias for the 21st century. Armonk (NY): Futura; 1998. p 473–489.
141. Gold MR, et al. Clinical experience with a dual-chamber implantable cardioverter defibrillator to treat atrial tachyarrhythmias. J Cardiovasc Electrophysiol 2001;12: 1247–1253.
142. Swerdlow CD, et al. Detection of atrial fibrillation and flutter by a dual-chamber implantable cardioverter-defibrillator. Circulation 2000;101:878–885.
143. KenKnight BH, Lang DJ, Scheiner A, Cooper RAS. Atrial defibrillation for implantable cardioverter-defibrillators: lead systems, waveforms, detection algorithms, and results. In: Singer I, Barold SS, Camm AJ, editors. Nonpharmacological Therapy of Arrhythmias for the 21st century. Armonk (NY): Futura; 1998. p 457–471.
144. Costa M, Moody GB, Henry I, Goldberger AL. PhysioNet: an NIH research resource for complex signals. J Electrocardiol 2003;36(Suppl) 139–144. Available at http://www.physionet.org.
145. American Heart Association ECG Database, Available from ECRI, 5200 Butler Pike, Plymouth Meeting, PA 19462 USA, http://www.ecri.org/.
146. Reek S, et al. Clinical efficacy of a wearable defibrillator in acutely terminating episodes of ventricular fibrillation using biphasic shocks. PACE 2003;26:2016–2022.
147. Feldman AM, et al. Use of a wearable defibrillator in terminating tachyarrhythmias in patients at high risk for sudden death: results of WEARIT/BIROAD. PACE 2004;27:4–9.

See also AMBULATORY MONITORING; DEFIBRILLATORS; ELECTROCARDIOGRAPHY, COMPUTERS IN; EXERCISE STRESS TESTING.

ARTERIAL TONOMETRY. See TONOMETRY, ARTERIAL.

ARTIFICIAL BLOOD. See BLOOD, ARTIFICIAL.

ARTIFICIAL HEART. See HEART, ARTIFICIAL.

ARTIFICIAL HEART VALVE. See HEART VALVE PROSTHESES.

ARTIFICIAL HIP JOINTS. See HIP JOINTS, ARTIFICIAL.

ARTIFICIAL LARYNX. See LARYNGEAL PROSTHETIC DEVICES.

ARTIFICIAL PANCREAS. See PANCREAS, ARTIFICIAL.

ARTERIES, ELASTIC PROPERTIES OF

KOZABURO HAYASHI
Okayama University of Science
Okayama, Japan

INTRODUCTION

The elastic properties of the arterial wall are very important because they are closely related to arterial physiology and pathology, especially via effects on blood flow and arterial mass transport. Furthermore, stresses and strains in the arterial wall are prerequisite for the understanding of the pathophysiology and mechanics of the cardiovascular system. Stresses and strains cannot be analyzed without exact knowledge of the arterial elasticity.

STRUCTURE OF ARETRIAL WALL AND BASIC CHARACTERISTICS

Arteries become smaller in diameter with increasing distance from the heart, depending on functional demands (1). In concert with this reduction in size, their structure, chemical composition, and wall thickness-inner diameter ratio gradually change in a way that leads to a progressive increase both in stiffness and in their ability to change their inner diameter in response to a variety of chemical and neurological control signals.

Arterial wall is inhomogeneous not only structurally, but also histologically. It is composed of three layers (intima, media, and adventitia), which are separated by elastic membranes. Because the media is much thicker than the other two layers and supports load induced by blood pressure, its mechanical properties represent the properties of arterial wall. The media is mainly composed of elastin, collagen, and cells (smooth muscle cell and fibroblast). Roughly speaking, elastin gives an artery its elasticity, while collagen resists tensile forces and gives the artery its burst strength. Smooth muscle cells contract or relax in response to mechanical, chemical, and the other stimuli, which alters the deformed configuration of arteries. The wall compositions vary at different locations depending on required functions. For example, collagen and smooth muscle increase and elastin decreases at more distal sites in conduit arteries; the ratio of collagen to elastin increases in more distally located arteries. Collagen and elastin are essentially similar proteins, but collagen is very much stronger and stiffer than elastin. Therefore, the change of arterial diameter developed by blood pressure pulsation depends on the arterial site; it is larger in more proximal arteries.

Like most biological soft tissues, arteries undergo large deformation when they are subjected to physiological loading, and their force-deformation and stress–strain relations are nonlinear partly because of the above-mentioned inhomogeneous structure and partly because of the nonlinear characteristics of each component itself. Since collagen is a long-chained high polymer, it is intrinsically anisotropic. Moreover, not only collagen and elastin fibers, but also cells, are oriented in tissues and organs in order so that their functions be most effective. Inevitably, the arterial wall is mechanically anisotropic like many other biological tissues. Biological soft tissues including arterial wall demonstrate opened hysteresis loops in their force–deformation and stress–strain curves, which means that those tissues are viscoelastic. In such materials, the stress state is not uniquely determined by current strain, but depends also on the history of deformation. When a viscoelastic tissue is elongated and maintained at some length, load does not stay at a specific level, but decreases rather rapidly at first and then gradually (relaxation). If some constant load is applied to the tissue, it is elongated with time rather rapidly at first and then gradually (creep). Viscoelastic materials generally show different stress–strain properties under different strain rates. It is true, and higher strain rates give higher stresses. However, such a strain rate effect is not so much in biological soft tissues like arteries, namely, their elastic properties are not more sensitive to strain rate. Therefore, it is not always necessary to consider viscoelasticity for arterial mechanics; it is very often enough to assume wall material to be elastic. Many biological soft tissues contain water of $> 70\%$. Therefore, they hardly change their volume even if load is applied, and they are almost incompressible. The incompressibility assumption is very important in the formulation of constitutive laws of soft tissues, because it imposes a constraint on the strains and they are not independent.

MEASUREMENT OF ARTERIAL ELASTICITY

In Vitro Tests

It is widely recognized that the mechanical properties of blood vessels do not change for up to 48 h if tissues are stored at $\sim 4\ ^{\circ}C$ (1). One of the basic methods for the determination of the mechanical properties of biological tissues is uniaxial tensile testing on excised specimens. In this test, an increasing force is steadily applied to the longitudinal direction of a specimen, and the resulting specimen deformation is measured, which gives relations between stress (force divided by specimen cross-sectional area) and strain (specimen elongation divided by reference specimen length). This *in vitro* test is simple but, nevertheless, provides us with basic and useful information on

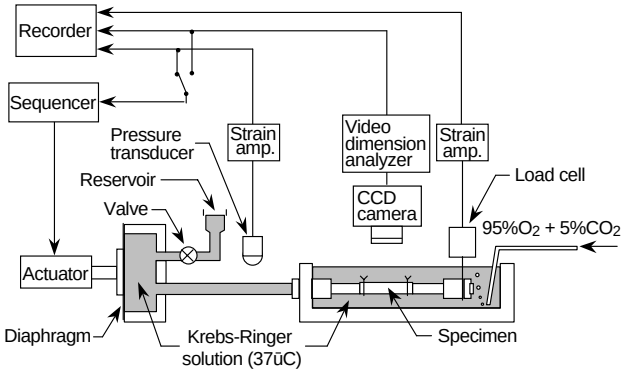


Figure 1. An *in vitro* experimental setup for the pressure–diameter–axial force test of a tubular arterial specimen. Internal pressure or outer diameter can be controlled with a feedback system (2).

the mechanical properties of tissues. Dumbbell-shape specimens, helically stripped specimens, and ring specimens are commonly used for arterial walls.

Under *in vivo* conditions, arteries are tethered to or constrained by perivascular connective tissues and side branches, and pressurized by blood from inside. These forces develop multiaxial stresses in the wall. For the determination of the mechanical characteristics of arteries under multiaxial conditions, biaxial tensile tests on flat specimens are utilized to simultaneously apply forces in the circumferential and longitudinal directions; however, the effect of wall radial stress is ignored in this case.

Although stress–strain data obtained from the above-mentioned uniaxial and biaxial tests on flat, strip, and ring specimens are often used to represent the elastic properties of arterial walls, the data obtained from pressure–diameter tests on the tubular segments of blood vessels are more important and realistic. An example of the test devices is shown in Fig. 1 (2). A tubular specimen is mounted in the bath filled with Krebs–Ringer solution, which is kept at 37 °C and aerated with 95% O_2 and 5% CO_2 gas mixture. Then, it is extend to the *in vivo* length to mimic the *in vivo* condition, because arteries inside the body are tethered to the surrounding tissues as mentioned above and, therefore, they are extended in the axial direction. A diaphragm-type actuator, which is controlled with a sequencer, is incorporated in the device for the application of internal pressure to the specimen. The internal pressure or specimen diameter can be controlled with the sequencer during pressure–diameter tests. The pressure is measured with a fluid-filled pressure transducer, while the outer diameter of the specimen is determined with a video dimension analyzer combined with a CCD camera. If the measurements of axial force are required in order to obtain pressure–diameter–axial force relations for the purpose of determining multiaxial constitutive laws, a load cell attached to one end of the vessel can be used.

In Vivo Measurements

It may be more realistic to obtain data from *in vivo* experiments under *in situ* conditions rather than to get data from *in vitro* biomechanical tests. As a result of recent progress in ultrasonic techniques, arterial diameter and even arterial wall thickness can be measured noninvasively with fairly good precision. These methods are being used not only for *in vivo* animal experiments, but also for clinical diagnosis of vascular diseases. It is true that the data obtained from these experiments and clinical cases are very useful, and provide important information concerning arterial mechanics. On the other hand, it is also true that many factors considerably affect the results obtained. These include physiological reactions to momentary changes in body and ambient conditions as well as the effects of anesthesia and respiration. In addition, since there has been some difficulty in applying the methods to small-diameter blood vessels, accurate measurements of vascular diameter and wall thickness with current techniques have been mostly limited to aortas and large arteries.

Before noninvasive ultrasonic techniques were developed, *in vivo* measurements of vascular diameter were invasively performed following surgical exposure of blood vessels, using strain gauge-mounted cantilevers, strain gauge-pasted calipers, and sonomicrometers. For example, a pair of miniature ultrasonic sensors may be used for the measurement of the outer diameter of a blood vessel (3). They are attached to the adventitial surface of a blood vessel so as to face each other across the vascular diameter. The diameter is determined from the transit time of the pulses between the two sensors. Similar sonomicrometers have been used for the measurement of arterial diameter not only in anesthetized, but also in conscious animals.

The noninvasive measurement of the elastic properties of arteries offers several significant advantages over invasive techniques. First, the nontraumatic character of the measurement guarantees a physiological state of the arterial wall, whereas such key functional elements of the wall as endothelium and smooth muscle might be affected in certain invasive measurement techniques. Second, it is of great clinical interest because it allows the monitoring of many outpatients and, therefore, it is well adapted for epidemiological or cross-sectional studies.

Noninvasive measurement of the arterial diameter can be done with ultrasonic echo-tracking techniques; recent improvements of the original technique have been proposed, which include digital tracking, prior inverse filtering, and coupling with B-mode imaging (1).

There exist no direct ways to measure pressure noninvasively in large central arteries, such as the aorta. Thus, regardless of the progress of ultrasonic and magnetic resonance imaging techniques which allow for the noninvasive measurement of vascular diameter, mechanical properties, such as compliance and elastic modulus cannot be derived from first principles. Therefore, primarily for clinical use, as an indirect, but noninvasive way of estimating the mechanical properties, the pulse wave velocity, c (see the next section), is often obtained from the measurements of pulsation at two distinct points along the vessel. One of the major drawbacks of this technique is low accuracy. The other one is that it yields a single value for the wave velocity. Because of the nonlinear elastic properties of the arterial wall, the pulse wave velocity sensitively changes depending on blood pressure. Therefore, the determination of a single value or a typical value of the arterial stiffness

estimated from the pulsation does not provide a full description of the mechanical properties of the arterial wall.

MATHEMATICAL EXPRESSION OF ARTERIAL ELASTICITY

Uniaxial Tensile Behavior

There are many tensile test data from arterial walls in humans and animals (4). Arterial walls exhibit nonlinear force-deformation or stress–strain behavior, having higher distensibility in the low force or stress range and losing it at higher force or stress. To represent strain in such biological soft tissues that deform largely and nonlinearly, we commonly use extension ratio, λ, which is defined by the ratio of the current length of a specimen (L) to its initial length (L_0). If we plot a stress/extension ratio curve as the slope of a stress/extension ratio curve versus stress, we can see that the relation is composed of one or two straight lines (1). Each line is described by

$$dT/d\lambda = BT + C \tag{1}$$

where T is Lagrangian stress defined by F/A_0 (F, force; A_0, cross-sectional area of an undeformed specimen), and B and C are constants. This is also expressed by

$$T = A[\exp B(\lambda - 1) - 1] \tag{2}$$

where A is equal to C/B. This type of exponential formulation is applicable to the description of the elastic behavior of many other biological soft tissues (5).

Pressure–Diameter Relations

For practical purposes, it is convenient to use a single parameter that expresses the arterial elasticity under living conditions. In particular, for noninvasive diagnosis in clinical medicine, material characterization should be simple, yet quantitative. For this purpose, several parameters have been proposed and commonly utilized (1). These include pressure–strain elastic modulus, E_p and vascular compliance, C_v. Pulse wave velocity, c, which was mentioned above, is also used to express elastic properties of the arterial wall. These parameters are described by

$$E_p = \Delta P/(\Delta D_o/D_o) \tag{3}$$

$$C_v = (\Delta V/V)/\Delta P \tag{4}$$

and

$$c^2 = (S/\rho)(\Delta P/\Delta S) = (V/\rho)(\Delta P/\Delta V) \tag{5}$$

where D_o, V, and S are the outer diameter, volume, and luminal area of a blood vessel at pressure P, respectively, and ΔD_o, ΔV, and ΔS are their increments for the pressure increment, ΔP. The parameter ρ is the density of the blood.

To calculate these parameters, we do not need to measure the wall thickness; for E_p and C_v, we need to know only pressure–diameter and pressure–volume data, respectively, at a specific pressure level. However, we should remember that these parameters express the stiffness or distensibility of a blood vessel. Therefore, they are structural parameters, and do not rigorously represent the inherent elastic properties of the wall material; in this sense, they are different from the elastic modulus which is explained below. In addition, these parameters are defined at specific pressures, and give different values at different pressure levels because the pressure–diameter relations of arteries are nonlinear.

To overcome this shortcoming, several functions have been proposed to mathematically describe pressure–diameter, pressure–volume, and pressure-luminal area data, and one or several parameters included in these equations have been used for the expression of the elastic characteristics of arteries. Among these functions, the following equation is one of the simplest and most reliable for the description of pressure-diameter relations of arteries in the physiological pressure range (6):

$$\ln(P/P_s) = \beta(D_o/D_s - 1) \tag{6}$$

where P_s is a standard pressure and D_s is the wall diameter at pressure P_s. A physiologically normal blood pressure like 100 mmHg (13.3 kPa) is recommended for the standard pressure, P_s. As an example, Fig. 2 shows the pressure–diameter relationships of a human femoral artery under normal and active conditions of vascular smooth muscle and the relations between the logarithm of pressure ratio, P/P_s, and distension ratio, D_o/D_s. Figure 2a demonstrates nonlinear behavior of the artery under both conditions, while Fig. 2b shows the close fit of the data to Eq. 6 over a rather wide pressure range. The coefficient, β, called the stiffness parameter, represents the structural stiffness of a vascular wall; it does not depend upon pressure. This parameter has been used for the evaluation of the stiffness of arteries not only in basic investigations, but also in clinical studies (1).

As can be seen from Fig. 2a, under the normal condition, arteries greatly increase the diameter with pressure under a low pressure range, say < 60 mmHg (8 kPa), and then gradually lose the distensibility at higher pressures. When vascular smooth muscle cells are activated by stimuli, arteries are contracted and their diameter decreases in a

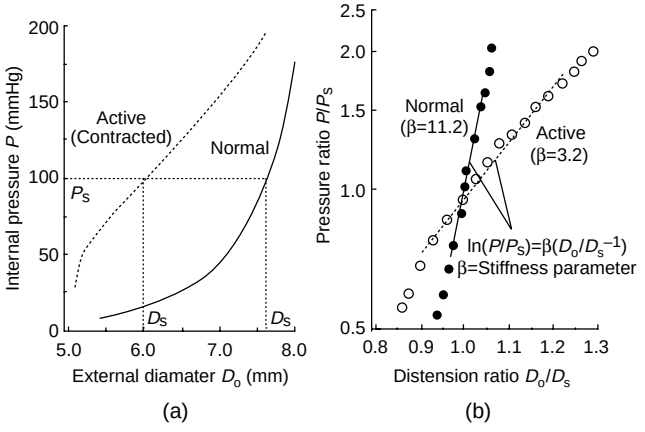


Figure 2. Pressure–diameter (a) and pressure ratio–distension ratio (b) relations of a human femoral artery under normal and active conditions (*in vitro* study) (1,7).

physiological pressure range and below the range [< 200 mmHg (26.6 kPa) in Fig. 2], and their pressure–diameter curves become greatly different from those observed under the normal condition.

To express the elastic properties of wall material, it is necessary to use a material parameter such as elastic modulus or Young's modulus, which is the slope of a linear stress–strain relation. For arterial walls that have nonlinear stress–strain relations, the following incremental elastic modulus has been often used for this purpose (8):

$$H_{\theta\theta} = 2D_i^2 D_o(\Delta P/\Delta D_o)/(D_o^2 - D_i^2) + 2PD_o^2/(D_o^2 - D_i^2) \quad (7)$$

where D_i is the internal diameter of a vessel. This equation was derived using the theory of small elastic deformation superposed on finite deformation in the case of a pressurized orthotropic cylindrical tube.

To calculate this modulus, it is necessary to know the thickness or internal diameter of a vessel. In *in vitro* experiments, we can calculate them from D_o, the internal and external diameters under no-load conditions measured after pressure–diameter testing, the *in vivo* axial extension ratio, and assuming the incompressibility of wall material. Noninvasive measurement of wall thickness or internal diameter on intact vessels has been rather difficult compared with the measurement of external diameter; however, it is now possible with high accuracy ultrasonic echo systems as mentioned above.

Constitutive Laws

Mathematical description of the mechanical behavior of a material in a general form is called a constitutive law or constitutive equation. We cannot perform any mechanical analyses without knowledge of constitutive laws of materials. Strain energy functions are commonly utilized for formulating constitutive laws of biological soft tissues that undergo large deformation (5). Let W be the strain energy per unit mass of a tissue, and ρ_0 be the density in the zero-stress state. Then, $\rho_0 W$ is the strain energy per unit volume of the tissue in the zero-stress state, and this is called the strain energy density function. Because arterial tissue is considered as an elastic solid, a strain energy function exists, and the strain energy W is a function solely of the Green strains:

$$W = W(E_{ij}) \quad (8)$$

where E_{ij} are the components of the Green strain tensor with respect to a local rectangular Cartesian coordinate system.

Under physiological conditions, arteries are subjected to axisymmetrical loads, and the axes of the principal stresses and strains coincide with the axes of mechanical orthotropy. Moreover, the condition of incompressibility is used to eliminate the radial strain E_{rr}, and therefore the strain energy function becomes a function of the circumferential and axial strains $E_{\theta\theta}$ and E_{zz}. Then, the constitutive equations for arteries are

$$\sigma_{\theta\theta} - \sigma_{rr} = (1 + 2E_{\theta\theta})\partial(\rho_0 W)/\partial E_{\theta\theta} \quad (9)$$

and

$$\sigma_{zz} - \sigma_{rr} = (1 + 2E_{zz})\partial(\rho_0 W)/\partial E_{zz} \quad (10)$$

where $\sigma_{\theta\theta}$, σ_{zz}, and σ_{rr} are Cauchy stresses in the circumferential, axial, and radial directions, respectively. Thus, we need to know the details of the strain energy function to describe stress–strain relations.

Three major equations have so far been proposed for the strain energy function of arterial wall. Vaishnav et al. (9) advocated the following equation:

$$\rho_0 W = (c/2)\exp(b_1 E_{rr}^2 + b_2 E_{\theta\theta}^2 + b_3 E_{zz}^2 + 2b_4 E_{rr}E_{\theta\theta} + 2b_5 E_{\theta\theta}E_{zz} + 2b_6 E_{zz}E_{rr}) \quad (11)$$

where $E_{\theta\theta}$ and E_{zz} are Green strains in the circumferential and axial directions, respectively, and A, B, and so on, are constants.

Chuong and Fung (10) proposed another form with an exponential function:

$$\rho_0 W = (c/2)\exp(b_1 E_{rr}^2 + b_2 E_{\theta\theta}^2 + b_3 E_{zz}^2 + 2b_4 E_{rr}E_{\theta\theta} + 2b_5 E_{\theta\theta}E_{zz} + 2b_6 E_{zz}E_{rr}) \quad (12)$$

where c, b_1, b_2, and so on, are material constants.

Later, Takamizawa and Hayashi (11) proposed a logarithmic form of the function described by

$$\rho_0 W = -C\ln(1 - a_{\theta\theta}E_{\theta\theta}^2/2 - a_{zz}E_{zz}^2/2 - a_{\theta z}E_{\theta\theta}E_{zz}) \quad (13)$$

where C, $a_{\theta\theta}$, a_{zz}, and $a_{\theta z}$ characterize the elastic properties of a material.

By using one of these strain energy equations or another type of equation for W in Eqs. 9 and 10, and applying the equations of equilibrium and boundary conditions, we determine the values of material constants. Although all of the proposed formulations describe quite well the elastic behavior of arterial walls, we prefer to reduce the number of constants included in the equations in order to handle them more easily, as well as to give physical meanings to the constants. For this reason, the logarithmic expression (Eq. 13) may be advantageous.

ELASTIC PROPERTIES OF NORMAL ARTERIES

Figure 3 shows β values of common carotid arteries, intracranial vertebral arteries, and coronary arteries obtained from autopsied human subjects of different ages (7). Note that arterial stiffness is much larger in the coronary arteries than in the other arteries, and also that intracranial vertebral arteries are considerably stiffer than extracranially located common carotid arteries. As can be seen from this figure, almost all the data obtained from normal human aortas and conduit arteries show that the structural stiffness of wall (e.g., E_p and β) increases with age rather gradually until the age of 40 years, and rapidly thereafter; on the other hand, the wall compliance (C_v) decreases with age. The stiffness of intracranial arteries like the intracranial vertebral artery progressively increases until 20 years, and then more slowly (6). There seems to be almost no age-related change in the human coronary artery (12).

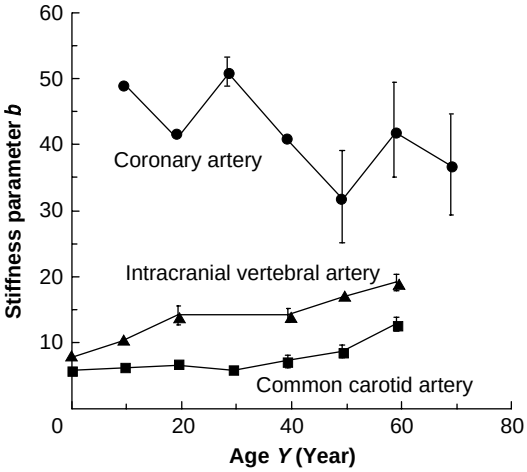


Figure 3. Age-related changes in the wall stiffness of human arteries. (*In vitro* studies, reproduced from Ref. 7.)

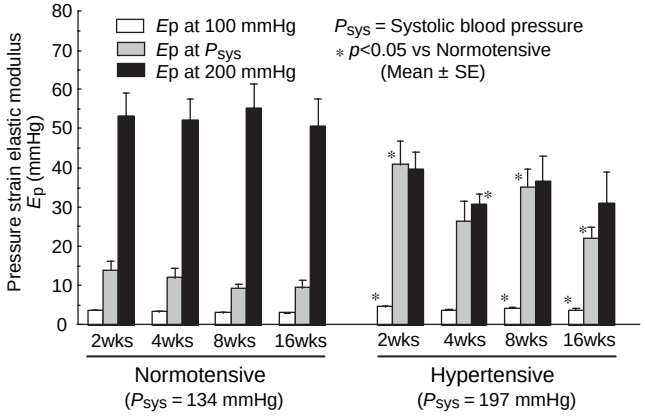


Figure 4. Structural stiffness of thoracic arotas in normotensive and hypertensive rats (7).

Most of the data obtained from arteries in animals indicated that the incremental elastic modulus or the slope of stress–strain curve increases with age, although there are several opposite data (13).

ELASTIC PROPERTIES OF DISEASED ARTERIES

Hypertension

Hypertension is recognized as one of the important risk factors for many cardiovascular diseases, including atherosclerosis and stroke. Elevated blood pressure exerts influences on the synthetic activity of vascular smooth muscle cells, and is believed to induce changes in structure and morphology of the arterial wall, its mechanical properties, and vascular contractility. It is therefore very important to understand arterial mechanics in hypertension. However, results from the extensive literature concerning the elastic properties of hypertensive arteries are contradictory and inconclusive (1,7,13,14). As mentioned above, when we analyze the reported data, we should remember that the values of such parameters as E_p (Eq. 3) and C_v (Eq. 4) are dependent on pressure. Without this consideration, comparisons between the results from different studies have little meaning.

Several studies have been performed to determine the pressure–diameter or pressure–volume relationships of aortas and arteries in hypertensive animals and humans. For example, comparison of hypertensive rats to normotensive controls has shown that at 100 mmHg (13.3 kPa) and also at the working pressure (systolic blood pressure before sacrifice, P_{sys}) of each group, the pressure strain elastic modulus, E_p of the thoracic aorta is greater in hypertensives than in normals; whereas at 200 mmHg (26.6 kPa) the E_p values in the hypertensive animals are slightly lower than those of the normals (Fig. 4) (7). These results do not depend on the duration of hypertension for 2–16 weeks.

With regard to the inherent elastic modulus of wall material calculated from pressure–diameter data, it has been shown that the incremental elastic moduli of the rat thoracic aorta ($H_{\theta\theta}$ in Eq. 7) at systolic blood pressure levels have significant correlations with blood pressure until 8 weeks after the induction of hypertension; at 16 weeks, however, the correlation disappears and the elastic modulus tends to be at the same level as that in control, normotensive rats (Fig. 5) (7). There are no significant differences in the incremental elastic modulus at

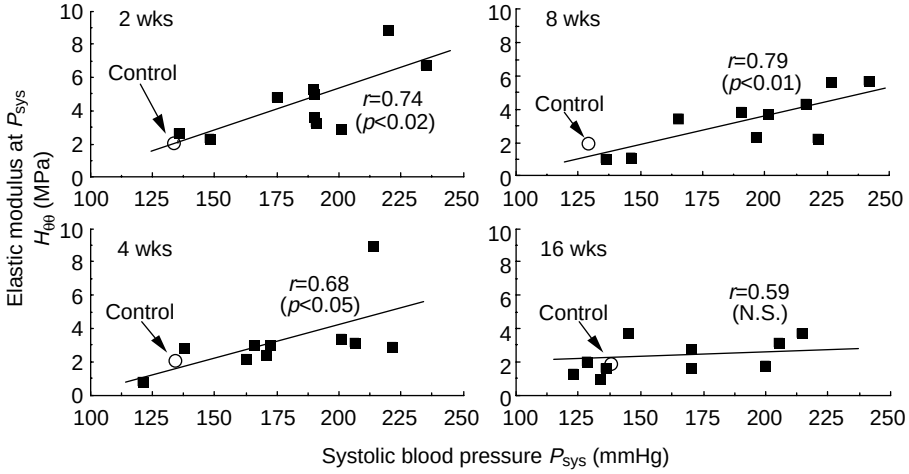


Figure 5. Incremental elastic modulus versus blood pressure in the rat aorta at 2, 4, 8, and 16 weeks after the treatment for hypertension (7).

100 mmHg (13.3 kPa) regardless of the period of hypertension. The aortic wall in hypertensive rats seems to restore the *in vivo* elastic properties to a normal level in 16 weeks due to the functional adaptation and remodeling of the wall.

In connection with the elastic properties of wall, many data have shown that the arterial wall is thickened by hypertension and hypertrophy occurs (14). Wall thickness critically depends on pressure level and, therefore, we have to pay attention to the pressure for the thickness measurement when we interpret the results. If the thickness of wall at the *in vivo* blood pressure level is used to calculate *in vivo* wall stress in the circumferential direction (hoop stress), the stress is independent of the degree of hypertension and is always maintained at a control, normal level even at 2 weeks after the induction of hypertension. This phenomenon is attributable to a functional adaptation and remodeling of the arterial wall (15).

Atherosclerosis

Effects of flow dynamics and wall shear stress on the initiation and development of atherosclerosis have been studied extensively. However, less attention has been paid to the mechanical properties of atherosclerotic wall tissue. Does atherosclerosis stiffen the arterial wall or increase the elastic modulus of wall? The results obtained have been conflicting and inconclusive, as shown in Table 1 (7). One of the reasons for this is that the structural stiffness of arterial wall and the elasticity of wall material have been confusingly used for the expression of the elastic properties of atherosclerotic wall. In this table, the elastic modulus represents the elasticity of wall material, which corresponds to the slope of stress–strain curve and is given by, for example, the incremental elastic modulus, $H_{\theta\theta}$; the stiffness is the structural stiffness expressed by, for example, the stiffness parameter, β.

Several studies have shown that the arterial wall is stiffened by the development of atherosclerosis. However, others have presented different results. Thus it has not been clear whether atherosclerosis increases the elastic modulus of arterial wall. We can see from Table 1 that atherosclerosis is mostly accompanied by wall thickening. This might be a reason why there are no data indicating a decrease in the structural stiffness associated with atherosclerosis. The structural stiffness is determined not only by the elastic modulus of wall material, but also by wall dimensions such as wall thickness.

A detailed and systematic study on the mechanical properties and morphology of atherosclerotic aortas in the rabbit has shown that the changes in the wall stiffness (β) and the elastic modulus ($H_{\theta\theta}$) are not always correlated

Table 1. Distributions of Reported Data of the Elasticity, Stiffness, and Thickness of Atherosclerotic Wall[a]

	No. of Data	Increase, %	No Change, %	Decrease, %
Elastic modulus	17	29	47	24
Stiffness	17	71	29	0
Wall thickness	12	67	25	8

[a]See Ref. 7

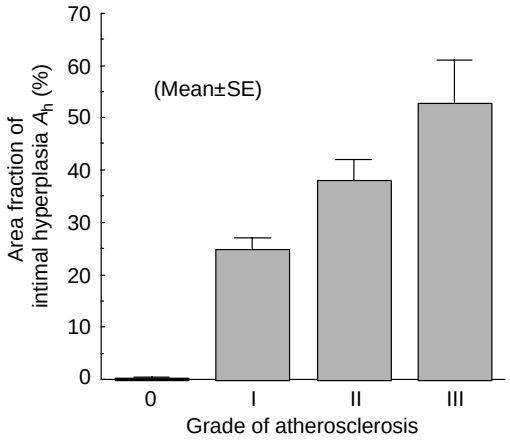


Figure 6. Area fraction of intimal hyperplasia in atherosclerotic thoracic aortas in the rabbit (7). Atherosclerosis was induced by the combination of denudation of endothelial cells and cholesterol diet.

with the time of cholesterol diet feeding (16). Thus, the grade of atherosclerosis was defined from the percent fraction of the luminal surface area stained with Sudan IV as well as from wall stiffening. The area fraction of intimal hyperplasia increases with the grade (Fig. 6). Likewise, wall thickness steadily increases with the progression

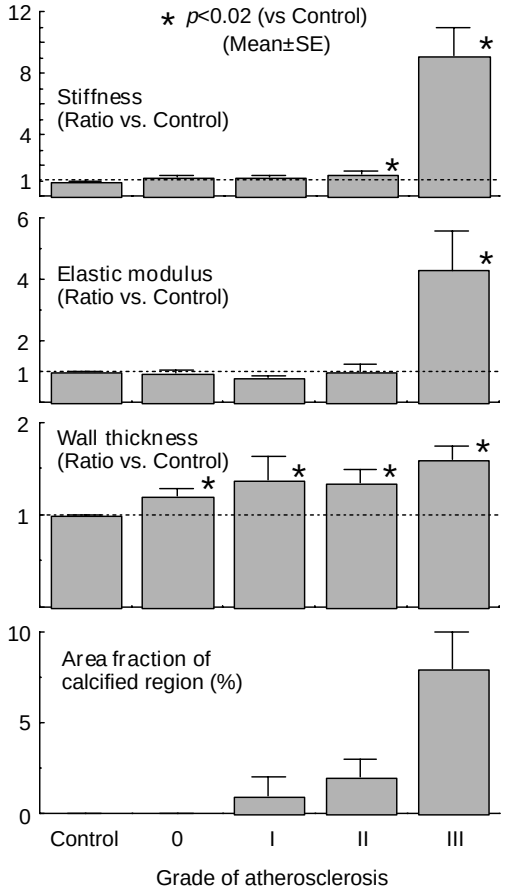


Figure 7. Elasticity, wall thickness, and calcification of athorosclerotic thoracic aortas in the rabbit (7).

of atherosclerosis (Fig. 7). However, the elastic modulus is not significantly different from the control artery until the highest grade of atherosclerosis. On the other hand, there appears a significant increase in the arterial stiffness in the grade II atherosclerosis, which is attributable to the wall thickening. Significantly increased calcification and intimal hyperplasia are observed in the wall of the grade III atherosclerosis. From these results, it is concluded that the progression of atherosclerosis induces wall thickening, followed by wall stiffening. However, even if atherosclerosis is advanced, there is essentially no change in the elastic modulus of wall material unless considerable calcification occurs in the wall. Calcified aortas have high elastic moduli. At the most advanced stage of atherosclerosis, the arterial wall has high structural and material stiffness due to calcification and wall hypertrophy (16).

BIBLIOGRAPHY

Cited References

1. Hayashi K, et al. Techniques in the Determination of the Mechanical Properties and Constitutive Laws of Arterial Walls. In: Leondes C, editor. Cardiovascular Techniques – Biomechanical Systems: Techniques and Applications Vol. II. Boca Raton (FL): CRC Press; 2001. p 6-1–6-61.
2. Hayashi K, Mori K, Miyazaki H. Biomechanical response of femoral vein to chronic elevation of blood pressure in rabbits. Am J Physiol 2003;284:H511–H518.
3. Hayashi K, Nakamura T. Implantable miniature ultrasonic sensors for the measurement of blood flow. Automedica 1989;12:53–62.
4. Abe H, Hayashi K, Sato M, editors. Data Book on Mechanical Properties of Living Cells, Tissues, and Organs. Tokyo: Springer-Verlag; 1996. p 25–125.
5. Fung YC. Biomechanics: Mechanical Properties of Living Tissues. New York: Springer-Verlag; 1993. p 242–320.
6. Hayashi K, et al. Stiffness and elastic behavior of human intracranial and extracranial arteries. J Biomech 1980;13: 175–184.
7. Hayashi K. Mechanical Properties of Soft Tissues and Arterial Walls. In: Holzapfel G, Ogden RW, editors. Biomechanics of Soft Tissue in Cardiovascular Systems. (CISM Courses and Lectures No. 441), Wien: Springer-Verlag; 2003. p 15–64.
8. Hudetz AG. Incremental elastic modulus for orthotropic incompressible arteries. J Biomech 1979;12:651–655.
9. Vaishnav RN, Young JT, Patel DJ. Distribution of stresses an strain energy density through the wall thickness in a canine aortic segment. Circ Res 1973;32:577–583.
10. Chuong CJ, Fung YC. Three-dimensional stress distribution in arteries. Trans ASME J Biomech Eng 1983;105:268–274.
11. Takamizawa K, Hayashi K. Strain energy density function and uniform strain hypothesis for arterial mechanics. J Biomech 1987;20:7–17.
12. Hayashi K, Igarashi Y, Takamizawa K. Mechanical properties and hemodynamics in coronary arteries. In: Kitamura K, Abe H, Sagawa K, editors. New Approaches in Cardiac Mechanics. New York: Gordon and Breach; 1986. p 285–294.
13. Hayashi K. Experimental approaches on measuring the mechanical properties and constitutive laws of arterial walls. Trans ASME J Biomech Eng 1993;115:481–488.
14. Humphrey JD. Cardiovascular Solid Mechanics: Cells, Tissues, and Organs. New York: Springer-Verlag; 2002. p 365–386.
15. Matsumoto T, Hayashi K. Response of arterial wall to hypertension and residual stress. In: Hayashi K, Kamiya A, Ono K, editors. Biomechanics: Functional Adaptation and Remodeling. Tokyo: Springer-Verlag; 1996. p 93–119.
16. Hayashi K, Ide K, Matsumoto T. Aortic walls in atherosclerotic rabbits -Mechanical study. Trans ASME J Biomech Eng 1994;116:284–293.

See also BLOOD PRESSURE, AUTOMATIC CONTROL OF; BLOOD RHEOLOGY; CUTANEOUS BLOOD FLOW, DOPPLER MEASUREMENT OF; HEMODYNAMICS; INTRAAORTIC BALLOON PUMP; TONOMETRY, ARTERIAL.

ASSISTIVE DEVICES FOR THE DISABLED.

See ENVIRONMENTAL CONTROL.

ATOMIC ABSORPTION SPECTROMETRY.

See FLAME ATOMIC EMISSION SPECTROMETRY AND ATOMIC ABSORPTION SPECTROMETRY.

AUDIOMETRY

THOMAS E. BORTON
BETTIE B. BORTON
Auburn University Montgomery
Montgomery, Alabama

JUDITH T. BLUMSACK
Disorders Auburn University
Auburn, Alabama

INTRODUCTION

Audiology is, literally, the science of hearing. In many countries around the world, audiology is a scientific discipline practiced by audiologists. According to the American Academy of Audiology, "an audiologist is a person who, by virtue of academic degree, clinical training, and license to practice and/or professional credential, is uniquely qualified to provide a comprehensive array of professional services related to the prevention of hearing loss and the audiologic identification, assessment, diagnosis, and treatment of persons with impairment of auditory and vestibular function, as well as the prevention of impairments associated with them. Audiologists serve in a number of roles including clinician, therapist, teacher, consultant, researcher and administrator" (1).

An important tool in the practice of audiology is audiometry, which is the measurement of hearing. In general, audiometry entails one or more procedures wherein precisely defined auditory stimuli are presented to the listener in order to elicit a measurable behavioral or physiologic response. Frequently, the term audiometry refers to procedures used in the assessment of an individual's threshold of hearing for sinusoidal (pure tones) or speech stimuli (2). So-called conventional audiometry is conducted with a calibrated piece of electronic instrumentation called an audiometer to deliver controlled auditory signals to a listener. Currently, however, an expanded definition of audiometry also includes procedures for measuring various physiological and behavioral responses to the presentation

of auditory stimuli, whether or not the response involves cognition. More sophisticated procedures and equipment are increasingly used to look beyond peripheral auditory structures in order to assess sound processing activity in the neuroauditory system.

Today, audiologists employ audiometric procedures and equipment to assess the function of the auditory system from external ear to brain cortex and serve as consultants to medical practitioners, education systems, the corporate and legal sectors, and government institutions such as the Department of Veterans Affairs. Audiologists also use audiometric procedures to identify and define auditory system function as a basis for nonmedical intervention with newborns, young children with auditory processing disorders, and adults who may require sophisticated amplification systems to develop or maintain their communication abilities and quality of life.

The purpose of this chapter is to acquaint the reader with the basic anatomy of the auditory system, describe some of the instrumentation and procedures currently used for audiometry, and briefly discuss the application of audiometric procedures for the assessment of hearing.

AUDIOMETRY AND ITS ORIGINS

Audiometry refers broadly to qualitative and quantitative measures of auditory function/dysfunction, often with an emphasis on the assessment of hearing loss. It is an important tool in the practice of audiology, a healthcare specialty concerned with the study of hearing, and the functional assessment, diagnosis, and (re)habilitation of hearing impairment. The profession sprang from otology and speech pathology in the 1920s, about the same time that instrumentation for audiometry was being developed. Audiometry grew rapidly in the 1940s when World War II veterans returned home with hearing impairment related to military service (3). Hearing evaluation, the provision of hearing aids, and auditory rehabilitation were pioneered in the Department of Veterans Affairs and, subsequently, universities began programs to educate and train audiologists for service to children and adults in clinics, hospitals, research laboratories and academic settings, and private practice. Audiometry is now a fundamental component of assessing and treating persons with hearing impairment.

Audiometers

Audiometers are electroacoustic instruments designed to meet internationally accepted audiometric performance standards for valid and reliable assessment of hearing sensitivity and auditory processing capability under controlled acoustic conditions. The audiometer was first described around the turn of the twentieth century (4) and was used mainly in laboratory research at the University of Iowa. A laboratory assistant at the university, C. C. Bunch, would later publish a classic textbook describing audiometric test results in patients with a variety of hearing disorders (5). The first commercial audiometer, called the Western Electric 1A, was developed in the early 1920s by the Bell Telephone Laboratories in the United States, and was described by Fowler and Wegel in 1922 (6). More than 20 years elapsed before the use of audiometers for hearing assessment was widely recognized (7), and it was not until the early 1950s (8) that audiometry became an accepted clinical practice. Since that time, electroacoustic instrumentation for audiometry has been described in standards written by scientists and experts designed to introduce uniformity and facilitate the international exchange of data and test results. The American National Standards Institute (ANSI), the International Standards Organization (ISO), and the International Electrotechnical Commission (IEC) are recognized bodies that have developed accepted standards for equipment used in audiometry and in psychophysical measurement, acoustics, and research. In the present day, audiometric procedures are routinely used throughout the world for identifying auditory dysfunction in newborns, assessing hearing disorders associated with ear disease, and monitoring the hearing of patients at risk for damage to the auditory system (e.g., because of exposure to hazardous noise, toxic substances).

Types of Audiometers

In the United States, ANSI (9) classifies audiometers according to several criteria, including the use for which they are designed, how they are operated, the signals they produce, their portability, and other factors.

In general, Type IV screening audiometers (designed to differentiate those with normal hearing sensitivity from those with hearing impairment) are of simpler design than those instruments used for in-depth diagnostic evaluation for medical purposes (Type I audiometers). There are automatic and computer-processor audiometers (Bekesy or self-recording types), extended high-frequency (Type HF), free-field equivalent audiometers (Type E), speech audiometers (Types A, B, and C depending on available features), and others for specific purposes.

Audiometers possess one or more oscillation networks to generate pure tones of differing frequencies, switching networks to interrupt and direct stimuli, and attenuating systems calibrated in decibels (dB) relative to audiometric zero. The intensity range for most attenuation networks usually approximates 100 dB, typically graduated in steps of 5 dB. The "zero dB" level represents normal hearing sensitivity across the test frequency range for young adults under favorable, noise-free laboratory conditions. Collection of these hearing level reference data from different countries began in the 1950s and 1960s. These reference levels have been accepted by internationally recognized standards organizations. Audiometers also include various types of output transducers for presentation of signals to the listener, including earphones, bone-conduction vibrators, and loudspeakers. As the audiometrically generated signals are affected substantially by the electroacoustic characteristics of these devices (e.g., frequency response), versatile audiometers have multiple calibrated output networks to facilitate switching between transducers depending on the clinical application of interest.

Figure 1 shows the general layout of a Type I diagnostic audiometer. Such instruments are required by standards governing them to produce a stable output at a range of

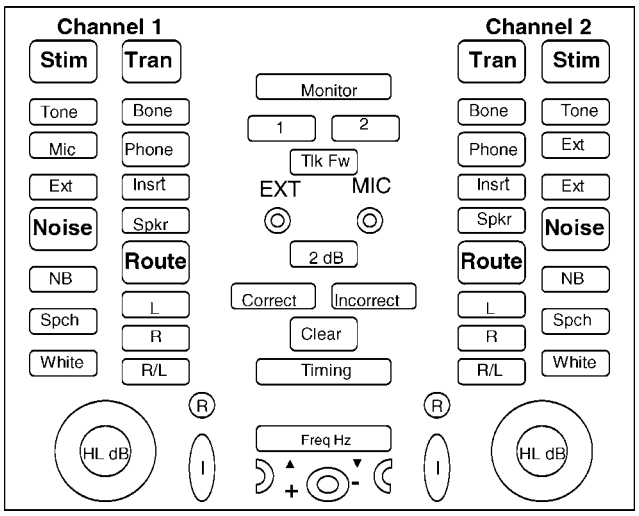


Figure 1. Typical diagnostic audiometer.

operating temperatures and humidity and meet a wide variety of electrical and other safety standards, in addition to precise electroacoustic standards for frequency, intensity, spectral purity, maximum output sound pressure level (SPL), and harmonic distortion. Type II audiometers have fewer required features and less flexibility, and Type IV audiometers have even more limitations.

Audiometric Calibration

To ensure that an audiometer is performing in accordance with the relevant standard, the instrument's electroacoustic characteristics are checked and adjusted as necessary, usually following a routine procedure. These calibration activities may be conducted at the manufacturing facility or an outside laboratory, but are most often accomplished on site at least annually. Calibration of speakers in a sound field is typically conducted on site because the unique acoustic characteristics of a specific field cannot easily be reproduced in a remote calibration facility.

Calibration of audiometers is routinely checked using instruments such as oscilloscopes, multimeters, spectrometers, and sound-level meters to verify frequency, intensity, and temporal characteristics of the equipment. Output transducers such as earphones and bone stimulators can be calibrated in two ways: (*1*) using "real ear" methods, involving individuals or groups of persons free from ear pathology and who meet other criteria, or (*2*) using hard-walled couplers (artificial ears) and pressure transducers specified by the relevant standard.

Audiometric Standards

Electroacoustic instrumentation for audiometry has been described in national and international standards written by scientists and experts designed to introduce uniformity and facilitate the international exchange of data and test results. ANSI, ISO, and IEC are recognized bodies that have developed accepted standards for equipment used in audiometry and acoustics. Some standards relate to equipment, others to audiometric procedures, and still others to the environment and conditions in which audiometry should be conducted (10).

The aim of standards for audiometric equipment and procedures is to assure precision of equipment functions to help ensure that test results can be interpreted meaningfully and reliably within and between clinics and laboratories using different equipment and personnel in various geographical locations. The results of audiometry often help to provide a basis for decisions regarding intervention strategies, such as medical or surgical intervention, referral for cochlear implantation, hearing aid selection and fitting, application of assistive listening devices, or selection of appropriate educational or vocational placement. As in any measurement scheme, audiometric test results can be no more precise than the function of the equipment and the procedures with which those measurements are made.

PURE TONE AUDIOMETRY

Psychophysical Methods

Audiometry may be conducted with a variety of methodologies depending on the goal of the procedure and subject variables such as age, mental status, and motivation. For example, the hearing sensitivity of very young children may be estimated by assessing the effects of auditory stimuli presented in a sound field on startle-type reflexes, level of arousal, and localization. Patients who are developmentally delayed may be taught with reinforcement to push a button upon presentation of a test stimulus. Children of preschool age may be taught to make a motor response to auditory test stimuli using play audiometric techniques.

In conventional audiometry, auditory stimuli are presented through special insert or supra-aural earphones, or a bone oscillator worn by the patient. When indicated, a sound field around the listener may be created by presenting stimuli through strategically placed loudspeakers. Most threshold audiometric tests in school-aged children and adults can be conducted using one of two psychophysical methods originally developed by Gustav Fechner: (*1*) the method of adjustment, or (*2*) the method of limits (11). In the method of adjustment, the intensity of an auditory stimulus is adjusted by the listener according to some criterion (just audible or just inaudible), usually across a range of continuously or discretely adjusted frequencies. Nobel Prize Laureate Georg von Bekesy initially introduced this methodology into the practice of audiometry in 1947 (12). With this approach, listeners heard sinusoidal stimuli that changed from lower to higher test frequencies, and adjusted the intensity of continuous and interrupted tones from "just inaudible" to "just audible". As shown in Fig. 2, this methodology yielded information about the listener's auditory threshold throughout the test frequency range. The relationship of threshold tracings for the pulsed and continuous stimuli added additional information about the site of lesion causing the hearing loss (13,14).

Later, it was found that the tracing patterns tracked by hearing-impaired patients at their most comfortable loudness levels, instead of their threshold levels, yielded additional useful diagnostic information (15). A myriad of

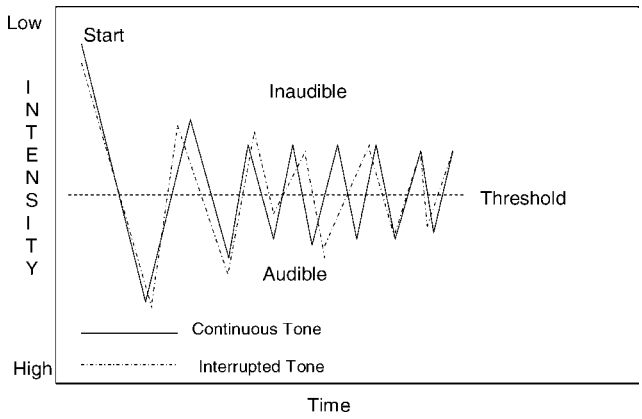


Figure 2. Bekesy audiometric tracing of continuous and interrupted tones around auditory threshold.

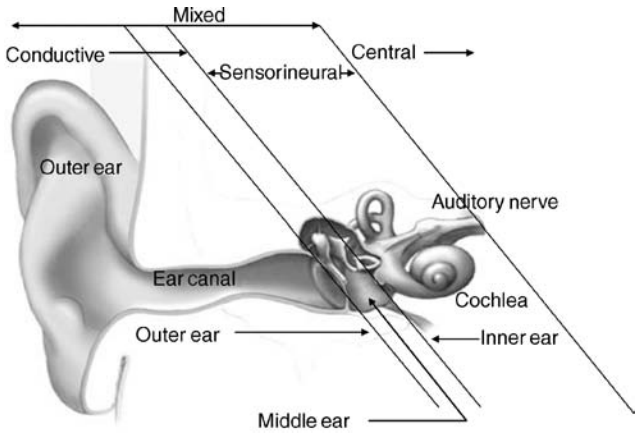


Figure 3. Anatomy of the peripheral auditory mechanism. Adapted from medical illustrations by NIH, Medical Arts & Photography Branch.

factors associated with this psychophysical measurement method, including the age of the listener, learning effects, and the length of time required to obtain stable test results, make these methods unsuitable for routine diagnostic purposes, especially in young children. Nevertheless, audiometers incorporating this methodology are manufactured to precise specifications (9) and are routinely used in hearing conservation programs to record the auditory sensitivity of large numbers of employees working in industrial or military settings.

In most clinical situations today, routine threshold audiometry is conducted using the method of limits. In this approach, the examiner adjusts the intensity of the auditory stimuli of various frequencies according to a predetermined schema, and the listener responds with a gross motor act (such as pushing a button or raising a hand) when the stimulus is detected. Although auditory threshold may be estimated using a variety of procedural variants (ascending, descending, mixed, adaptive), research has established (16) that an ascending approach in which tonal stimuli are presented to the listener from inaudible intensity to a just audible level is a valid and reliable approach for cooperative and motivated listeners, and the technique most parsimonious with clinical time and effort. In this approach, tonal stimuli are presented at intensity levels below the listener's threshold of audibility and raised in increments until a response is obtained. At this point, the intensity is lowered below the response level and increased incrementally until a response is obtained. When the method of limits is used, auditory threshold is typically defined as the lowest intensity level that elicits a reliable response from the patient on approximately 50% or more of these "ascending" trials.

Sound Pathways of the Auditory System

The fundamental anatomy of the ear is depicted in Fig. 3. Sound enters the auditory mechanism by two main routes, air conduction and bone conduction. Most speech and other sounds in the ambient environment enter the ear by air conduction. The outer ear collects and funnels sound waves into the ear canal, provides a small amount of amplification to auditory signals, and conducts sound to the tympanic membrane (eardrum). Acoustic energy strikes the tympanic membrane, where it is converted to mechanical energy in the form of vibrations to be conducted by small bones across the middle ear space to the inner ear. These mechanical vibrations are then converted to hydraulic energy in the fluid-filled inner ear (cochlea). This hydrodynamic form of energy results in traveling waves on cochlear membranous tissues. Small sensory hair cells are triggered by these waves to release neurotransmitters, resulting in the production of neural action potentials that are conducted through the auditory nerve (N. VIII) via central auditory structures in the brainstem to the auditory cortex of the brain, where sound is experienced.

Disorders affecting different sections of the ear depicted in Fig. 3 produce different types of hearing impairment. The outer ear, external auditory canal, and ossicles of the middle ear are collectively considered as the conductive system of the ear, and disorders affecting these structures produce a conductive loss of hearing. For example, perforation of the tympanic membrane, presence of fluid (effusion) in the middle ear due to infection, and the disarticulation of one or more bones in the middle ear all produce conductive hearing loss. This type of hearing loss is characterized by attenuation of sounds transmitted to the inner ear, and medical/surgical treatment often fully restores hearing. In a small percentage of cases, the conductive disorder may be permanent, but the use of a hearing aid or other amplification device can deliver an adequate signal to the inner ear that usually permits excellent auditory communication.

The inner ear and auditory nerve comprise the sensorineural mechanism of the ear, and a disorder of this apparatus often results in a permanent sensorineural hearing loss. Sensorineural disorders impair both perceived sound audibility and sound quality typically because of impaired frequency selectivity and other effects. Thus, in sensorineural-type impairments, sounds become difficult to detect, and they are also unclear, leading to poor understanding of speech. In some cases, conductive and sensorineural disorders simultaneously co-exist to produce

a mixed-type hearing impairment. Listeners with this disorder experience the effects of conductive and sensorineural deficits in combination.

Finally, the central auditory system begins at the point the auditory nerve enters the brainstem, and comprises the central nerve tracts and nuclear centers from the lower brainstem to the auditory cortex of the brain. Disorders of the central auditory nervous system produce deficits in the ability to adequately process auditory signals transmitted from the outer, middle, and inner ears. The resulting hearing impairment is characterized not by a loss of sensitivity to sound, but rather difficulties in identifying, decoding, and analyzing auditory signals, especially in difficult listening environments with background noise present. Auditory processing disorders require sophisticated test paradigms to identify and diagnose.

The Audiogram

The results of basic audiometry may be displayed in numeric form or on a graph called an audiogram, as shown in Fig. 4. As can be seen, frequency in hertz (Hz) is depicted on the abscissa, and hearing level (HL) in dB is displayed on the ordinate. Although the normal human ear can detect frequencies below 100 Hz and as high as 20,000 Hz, the audible frequency range most important for human communication lies between 125 and 8000 Hz, and the audiogram usually depicts this more restricted range. For special diagnostic purposes, extended high frequency audiometers produce stimuli between 8000 and 20,000 Hz, but specialty audiometers and earphones must be used to obtain thresholds at these frequencies. A few commercially available audiometers produce sound pressure levels as high as 120 dB HL, but such levels are potentially hazardous to the human ear and hearing thresholds poorer than 110 dB do not represent "useful" hearing for purposes of communication.

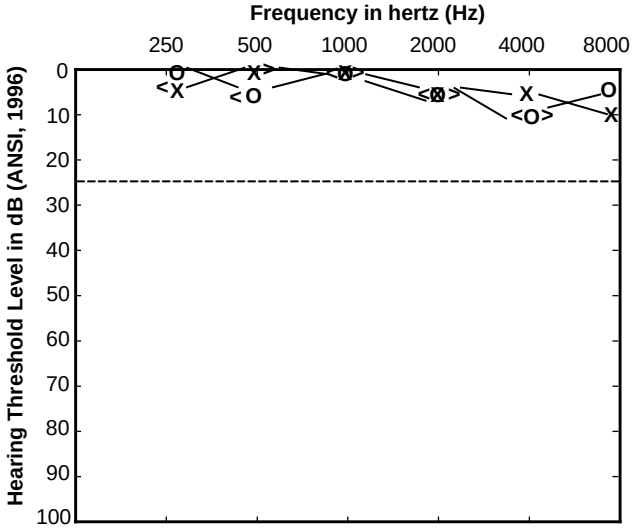


Figure 4. Graphic audiogram for a normal hearing listener. Bone conduction, right ear = <; Bone conduction, left ear = >; Air conduction, right ear = O; Air conduction, left ear = X.

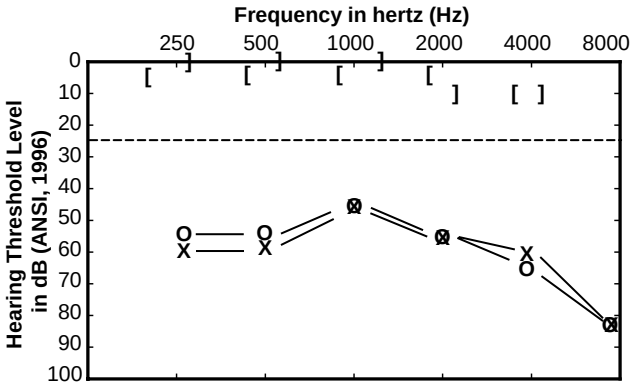


Figure 5. Graphic audiogram for a listener with conductive hearing loss. Bone conduction, right ear = [; Bone conduction, left ear =]; Air conduction, right ear = O; Air conduction, left ear = X.

The dashed line across the audiogram in Fig. 4 at 25 dB HL represents a common depiction of the boundary between normal hearing levels and the region of hearing loss (below the line) in adults. The recorded findings on this audiogram represent normal test results from an individual with no measurable loss of hearing sensitivity.

Figure 5 displays test results for a listener with a middle ear disorder in both ears and a bilateral conductive loss of hearing, which is moderate in degree, and similar in each ear. Bone conduction responses for the two ears are within normal limits (between 0 and 25 dB HL), suggesting normal sensitivity of the inner ear and auditory nerve, while air conduction thresholds are depressed below normal, suggesting obstruction of the air conduction pathway to the inner ear. Thus, conductive hearing losses are characterized on the audiogram by normal bone conduction responses and depressed air conduction responses. In

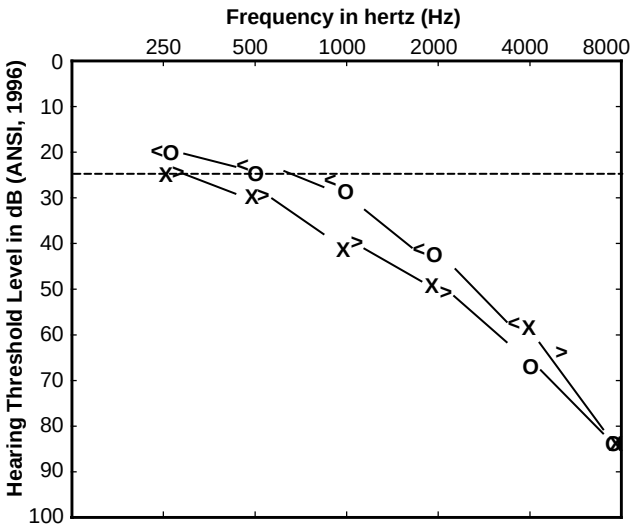


Figure 6. Graphic audiogram. Bone conduction, right ear = <; Bone conduction, left ear = >; Air conduction, right ear = O; Air conduction, left ear = X.

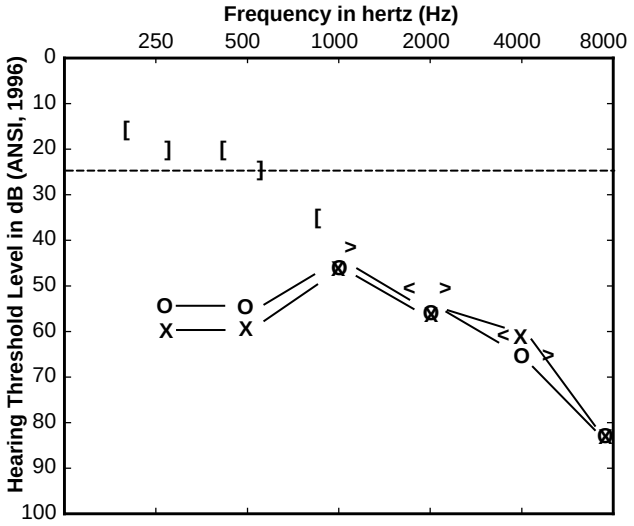


Figure 7. Graphic audiogram for a listener with mixed hearing loss. Masked bone conduction, right ear = [; Masked bone conduction, left ear =]; Unmasked bone conduction, right ear = <; Unmaksked bone conduction, left ear = >; Air conduction, right ear = O; Air conduction, left ear = X.

sensorineural-type hearing losses, air conduction and bone conduction responses in each ear are equally depressed on the audiogram. Figure 6 shows a high frequency loss of hearing in both ears, falling in pattern and sensorineural in type. Air and bone conduction hearing sensitivity is similar in both ears, suggesting that the cause of the hearing loss is not in the conductive mechanism (outer and middle ears).

The audiogram shown in Fig. 7 depicts a mixed-type loss of hearing in both ears. The gap between air and bone conduction thresholds in the two ears at the lower frequencies suggests a conductive disorder affecting the outer or middle ears. However, at frequencies above 500 Hz, hearing sensitivity via both air and bone conduction pathways in the two ears is nearly identical, which points to a disorder affecting the inner ear or auditory nerve.

In summary, an audiogram displays the results of basic audiometry in a stylized "shorthand", so that the hearing impairment can be readily characterized according to type of loss, degree of deficit, configuration (shape) of loss, and the degree of symmetry between the two ears. Such findings constitute the basis for first-order description of a listener's hearing sensitivity across the audible frequency range and provide important clues about the cause of hearing loss, the effects of the impairment on auditory communication ability, and the prognosis for treatment and rehabilitation.

SPEECH AUDIOMETRY

The first attempts to categorize hearing impairment on the basis of tests using speech stimuli were made in the early 1800s, when sounds were ranked according to their intensity and used to estimate the degree of hearing loss (17). Throughout the 1800s, refinements were introduced in methodologies for using speech stimuli to evaluate hearing. These improvements included the control of word intensities by varying distance between speaker and listener, the introduction of whispered speech to reduce differences in audibility between words, recording speech stimuli on the phonograph devised by Edison in 1877, and standardizing words lists in English and other languages (17). Most of the early research on speech perception focused on the sensitivity of the auditory system to speech, but progress in this area accelerated in the early 1900s because of work at the Bell Telephone Laboratories centered on the discrimination of speech sounds from each other. This emphasis led to the development of modern materials for assessing speech recognition at the Harvard Psychoacoustic Laboratories (18), which have been refined and augmented since that time.

Although pure tone audiometry provides important information about hearing sensitivity, as well as the degree, configuration, and type of hearing loss in each ear, it provides little information about a listener's auditory communication status and the ability to hear and understand speech in quiet as well as difficult listening situations. Attempts to predict speech recognition ability from the pure tone audiogram, even with normal hearing listeners, have met with only partial success, and the task is particularly complicated when listeners have a hearing impairment.

Instrumentation

Speech audiometry is conducted in the "speech mode" setting of a clinical audiometer. Speech stimuli are presented through the same types of transducers as those used for pure tone audiometry. Live speech stimuli via microphone and monitored with a VU meter can be used for speech audiometry, or recorded speech materials can be presented by CD or tape and routed through the audiometer to either one ear or both ears simultaneously by earphone or loudspeaker. Recorded speech materials typically include a calibration tone, and the input level is adjusted for individual recordings to a specified intensity level. Many different speech audiometric tests have been developed, and most currently in use are available in recorded form. Monitored live-voice presentation enables greater flexibility, but recorded speech materials enhance consistency across test conditions and avoid performance differences related to talker speech and vocal eccentricities.

In general, speech audiometry is conducted with the examiner in one room and the listener in another. With this arrangement, the examiner is able to observe the listener and maintain easy communication through microphones in both rooms, but the speech stimuli can be presented under carefully controlled conditions.

Speech Recognition Threshold

Speech recognition threshold (SRT) testing typically entails presentation of spondees (two-syllable, compound words), spoken with equal stress on each syllable (e.g., baseball, toothbrush, airplane). The use of these words for audiometric purposes has been investigated extensively, especially with respect to similarity in audibility (19).

Audiologists now generally select stimulus words from a list of commonly accepted spondees, and the words are presented at varying intensities using protocols similar to those used for pure tone audiometry. The speech recognition threshold (SRT) is the lowest intensity level at which the patient correctly responds to (repeats, writes down, points to) approximately 50% of the words, with the goal of determining the threshold of hearing for speech. The relationship between thresholds for speech and pure tone was identified in the early part of the twentieth century (20) and later described in detail (21,22). For purposes of clinical speech audiometry, speech recognition thresholds are expected to be within ± 6 dB of the average of the patient's pure tone air conduction thresholds at 500, 1000, and 2000. However, if the pure tone air conduction thresholds slope steeply, the speech recognition threshold is expected to agree with the average of the two best pure tone thresholds in the range of 500–2000 Hz.

The expected agreement between pure tone thresholds and speech recognition thresholds enables audiologists to use the SRT as a cross-check of pure tone air conduction threshold values. Disagreement between SRTs and pure tone threshold averages occurs for a variety of reasons. For example, poor agreement may exist between pure tone thresholds and SRTs in each ear if the patient misunderstands instructions regarding the test procedure for pure tone audiometry, or if the patient attempts to deceive the audiologist regarding actual hearing sensitivity.

SRTs can also be used to estimate/predict pure tone air conduction thresholds in the so-called speech frequency range of 500–2000 Hz in patients who are difficult to test with pure tones. Young children, for example, may reliably repeat or point to pictures of spondees (baseball, toothbrush) while exhibiting inconsistent responses to more abstract pure tones. Speech recognition thresholds have also been used as a basis for predetermining the presentation level for suprathreshold speech stimuli.

Speech Detection Threshold

Whereas the SRT represents the least intensity at which 50% of the speech stimuli presented to the listener can be recognized, the Speech Detection Threshold (SDT), sometimes called the Speech Awareness Threshold (SAT), represents the lowest intensity at which the patient exhibits an awareness of the presence of speech stimuli. If thresholds for spondaic words cannot be established, because of language impairment or other limitations such as young age or inability to speak because of injury, the SDT may represent a useful estimate of the level at which the patient indicates awareness of the presence of speech. In this type of speech threshold testing, the patient is not required to repeat the speech stimulus, which may be just a simple word or nonsense sound, but, instead, the patient simply responds with a hand movement or other gesture to indicate that a sound was detected. The SDT is obtained with protocols similar to those used for speech recognition measurement, and it is expected to be approximately 7–9 dB less intense than the value that would be obtained for the SRT (23,24).

Suprathreshold Speech Audiometry

In suprathreshold speech audiometry, speech stimuli (live-voice or recorded) are presented at levels well above the speech threshold in order to assess the listener's ability to understand speech. Depending on the purpose of the evaluation, the stimuli may be presented in quiet or in the presence of noise (e.g., speech babble, speech-spectrum noise), and the stimuli may be single nonsense syllables, monosyllabic words, nonsense sentences, or sentences. For some purposes, the stimuli are intentionally degraded by filtering or mixing them with noise, and depending on the purpose of suprathreshold evaluation, stimuli may be presented to one ear only (monaurally) or to both ears (binaurally). When stimuli are presented binaurally (both ears simultaneously), they may be identical (diotic) or different (dichotic). Stimuli may be presented at a specified level greater than speech recognition threshold or at varying intensity levels to establish a performance-intensity function. In suprathreshold testing, patient performance is often characterized in terms of percent correct, and standardized norms are used to interpret results.

Purposes for assessment of speech understanding include assessing auditory communication impairment, evaluating effectiveness of a hearing aid fitting, facilitating a comparison between hearing aids, and detecting possible VIIIth nerve or central auditory processing disorder. Suprathreshold stimuli may also be used to determine most comfortable and uncomfortable listening levels for purposes related to hearing aid fitting.

ELECTROPHYSIOLOGIC AUDIOMETRY

Auditory Evoked Potentials—Introduction

The electrophysiological response of the auditory system is often used by audiologists to evaluate auditory function. The techniques are derived from electroencephalography (EEG), which is the measurement of ongoing neural activity and has long been used to monitor brain function. The EEG can be recorded with surface electrodes attached to the scalp and connected to instrumentation that amplifies and records neural activity. Embedded in ongoing EEG activity is the brain's specific response to sensory stimulation. Auditory nervous system responses can be intentionally evoked with an auditory stimulus (such as an acoustic click) presented via an earphone (or other transducer) coupled to the ear. Neural responses that are time-linked to the stimulus can be recorded and differentiated from background EEG activity and other electrical noise sources (e.g., muscle artifact, 60 Hz electrical line noise).

Auditory Evoked Potentials—Clinical Applications

Auditory evoked response recording is an important tool for estimation of auditory sensitivity, particularly when conventional audiometry cannot be used. Evoked auditory potentials are also used routinely to assess the integrity of the auditory system (e.g., in tumor detection, auditory processing assessment, intra-operative monitoring),

but the following discussion will focus on threshold estimation/prediction.

Auditory evoked responses are used in place of conventional audiometry primarily in (*1*) infant hearing screening and assessment, (*2*) auditory evaluation of noncooperative children and adults, and (3) threshold estimation for people whose neurological status precludes use of conventional techniques. Although evoked potentials are not true measures of hearing, evoked potentials can be used in conjunction with other tests and information to estimate or predict hearing sensitivity. The capacity to make such estimates has important implications for early identification and rehabilitation of hearing impairment in newborns and young children, provision of auditory rehabilitation to people who have neurological problems, and even evaluation of nonorganic hearing impairment.

Historical Perspective

Early work indicated that ongoing EEG activity can be modified by sensory input (25). In order for a response specific to sensory stimulation to be observed, however, it was necessary to develop techniques to extract the sensory response from the ongoing EEG voltages. One important extraction technique that was developed involved algebraic summation (often called averaging) of responses that are linked in time to the sensory stimulus (26). When a bioelectric potential that is time-locked to a stimulus is recorded repeatedly and added to itself, the amplitude of the observed response will gradually increase with each stimulus repetition. In contrast, as EEG voltages during the same recording period are random, EEG voltages, when repeatedly summed, will gradually diminish or average out. Signal averaging was a critical advancement toward the clinical use of auditory evoked potentials. Other developments followed, and clinical applications of auditory potentials have now been investigated extensively. Measurement and assessment of evoked potentials are currently standard components of audiological practice.

Instrumentation

Many systems for recording auditory evoked potentials are now commercially available and are used widely. Components of the recording equipment typically include a stimulus generator capable of generating a variety of stimuli (e.g., clicks, tone bursts, tones), an attenuator, transducers for stimulus presentation (e.g., insert earphones, standard earphones, bone oscillator), surface electrodes, a differential amplifier, amplifier, filters, analog-to-digital converter, a signal averaging computer, data storage, display monitor, and printer. A simplified schematic diagram of a typical instrument is shown in Fig. 8.

In preparation for a typical single-channel recording, three electrodes are placed on the scalp. The electrodes often are called noninverting, inverting, and ground, but other terminology may be used (e.g., positive/negative or active/passive). A typical electrode montage is shown in Fig. 8, but electrode placements may vary depending on the potentials being recorded and the judgment of the clinician. Unwanted electrical or physiologic noise that may distort or obscure features of the response is reduced

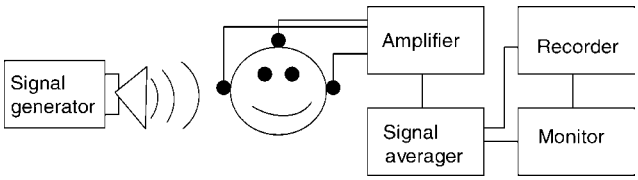


Figure 8. Simple block diagram of an auditory evoked response audiometry system.

by the use of differential amplifiers with high common mode rejection ratios and filters. It is important to note that electrode placement, stimulus polarity, stimulus presentation rate, number of signal presentations, signal repetition rate, filter characteristics, stimulus characteristics, and sampling rate during analog-to-digital conversion all affect the recording, and so must be controlled by the clinician.

Classification of Evoked Auditory Potentials

After the onset of an auditory stimulus, neural activity in the form of a sequence of waveforms can be recorded. The amount of time between the onset of the stimulus and the occurrence of a designated peak or trough in the waveform is called the latency. The latency of some auditory evoked potentials can be as short as a few thousandths of a second, and other latencies can be 400 ms or longer. Auditory evoked potentials are often classified on the basis of their latencies. For example, a system of classification can divide the auditory evoked potentials into "early" [$<$ 15 ms (e.g., electrocochleogram and auditory brainstem response)], "middle" [15–80 ms (e.g., Pa, Nb, and Pb)], and "late" [$>$ 80 ms (e.g., P300)] categories. Various classification systems based on latencies have been described, and other forms of classification systems based on the neural sites presumed to be generating the potentials (e.g., brainstem, cortex) are also sometimes used.

It is important to note that recording most bioelectric potentials requires only passive cooperation from the patient, but for some electrical potentials originating in the cortex of the brain, patients must provide active, cognitive participation. In addition, certain potentials are affected by level of consciousness. These factors, combined with the purpose of the evaluation, are important in the selection of the waveforms to be recorded.

Auditory Threshold Estimation/Prediction with AEPs

Auditory threshold estimations/predictions have been made on the basis of early, middle, and late potentials, but the evoked potentials most widely used for this purpose are those recorded within the first 10–15 ms after stimulus onset, particularly the so-called auditory brainstem response (ABR). An ABR evoked by a click consists of 5–7 peaks that normally appear in this time frame (27–29). Typical responses are shown in Fig. 9. The figure depicts three complete ABRs, and each represents the algebraic average of 2048 responses to a train of acoustic transient stimuli. The ABR is said to be time-locked such that each of the prominent peaks occurs in the normal listener at

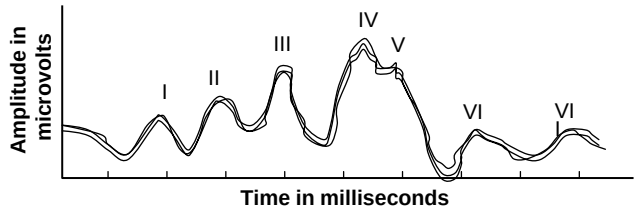


Figure 9. Normal auditory brainstem response; three complete responses.

predictable time periods after stimulation. Reliability is a hallmark of the ABR, and helps assure the audiologist that a valid estimation of conduction time through auditory brainstem structures has been made. A brief, gated, square-wave signal (click) stimulus is often used to generate the response, and stimulus intensity is reduced until the amplitude of the most robust peak (Wave V) is indistinguishable from the baseline voltage.

The response amplitude and latency (which lengthens as stimulus intensity decreases) are used to estimate behavioral auditory thresholds. In some equipment arrangements, computer software is used to statistically analyze the potentials for threshold determination purposes. ABR thresholds obtained with click stimuli correlate highly with behavioral thresholds at 2000–4000 Hz when hearing sensitivity ranges from normal to the severe range hearing impairment. Click stimuli are commonly used in clinical situations because their transient characteristics can excite many neurons synchronously, and thus a large amplitude response is evoked. However, variability limits the usefulness of click-evoked thresholds for prediction/estimation of auditory sensitivity of any particular patient (30), and the frequency specificity desired for audiometric purposes may not be obtained. As a result, gated tone bursts of differing frequencies are often used to estimate hearing sensitivity across the frequency range. These tonal stimuli can be embedded in bursts of noise to sharpen the frequency sensitivity and specificity of the test procedure.

In recent years, another evoked potential technique similar to the ABR has been developed to improve frequency specificity in threshold estimation while maintaining good neural synchronization. This technique, the auditory steady-state response (ASSR), uses rapidly (amplitude or frequency) modulated pure tone carrier stimuli (see Fig. 10). Evidence suggests that the ASSR is particularly useful when hearing sensitivity is severely impaired because high intensity stimuli can be used

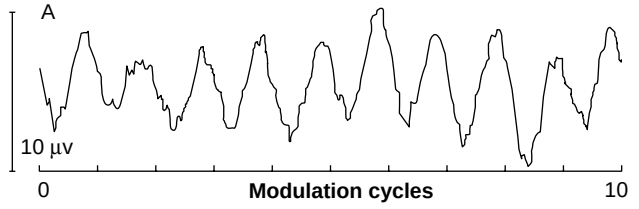


Figure 10. Auditory steady-state response.

(31). Research on the ASSR is ongoing, and this technique currently is considered to be a complement to click and tonal ABR in threshold estimation/prediction.

ACOUSTIC IMMITTANCE MEASUREMENT

Introduction

One procedure that helps audiologists interpret the results of conventional audiometry and other audiological tests is a measure of the ease with which energy can flow through the ear. Heaviside (1850–1925) coined the term impedance, as applied to electrical circuitry, and these principles were later applied in the United States during 1920s to acoustical systems (32). Mechanical impedance-measuring devices were initially designed for laboratory use, but electroacoustic measuring instruments were introduced for clinical use in the late 1950s (33). As acoustic impedance is difficult and expensive to measure accurately, measuring instruments using units of acoustic admittance are now widely used. The term used to describe measures incorporating the principles of both acoustic impedance and its reciprocal (acoustic admittance) is acoustic immittance. Modern instrumentation permits an estimate of ear canal and middle ear acoustic immittance (including resistive and reactive components).

Instrumentation

Commonly available immittance measuring devices (see Fig. 11) employ a probe-tone delivered to the tympanic membrane through the external ear canal. Sinusoids of differing frequencies are presented through a tube encased in a soft probe fitted snugly in the ear canal. The probe also contains a microphone and tubing connected to an air pump so that air pressure in the external ear canal can be varied from -400 to $+400$ mmH_2O.

Immittance devices also typically have a signal generator and transducers that can be used to deliver high intensity tones at various frequencies for the purpose of acoustic reflex testing. The American National Standards Institute (ANSI) has published a standard (S3.39-1987) for immittance instruments (34).

Immittance Measurement Procedures

As mentioned earlier in this chapter, the middle ear transduces acoustic energy into mechanical form. The transfer function of the middle ear can be estimated by measuring acoustic immittance at the plane of the tympanic membrane. These measures are often considered (in conjunction with the results of other audiological tests) in determining the site of lesion of an auditory disorder.

Static Acoustic Immittance

The acoustic immittance of the middle ear system is usually estimated by subtracting the acoustic immittance of the ear canal. This value is termed the compensated static acoustic immittance and is measured in acoustic mhos (reciprocal of acoustic ohms). The peak compensated

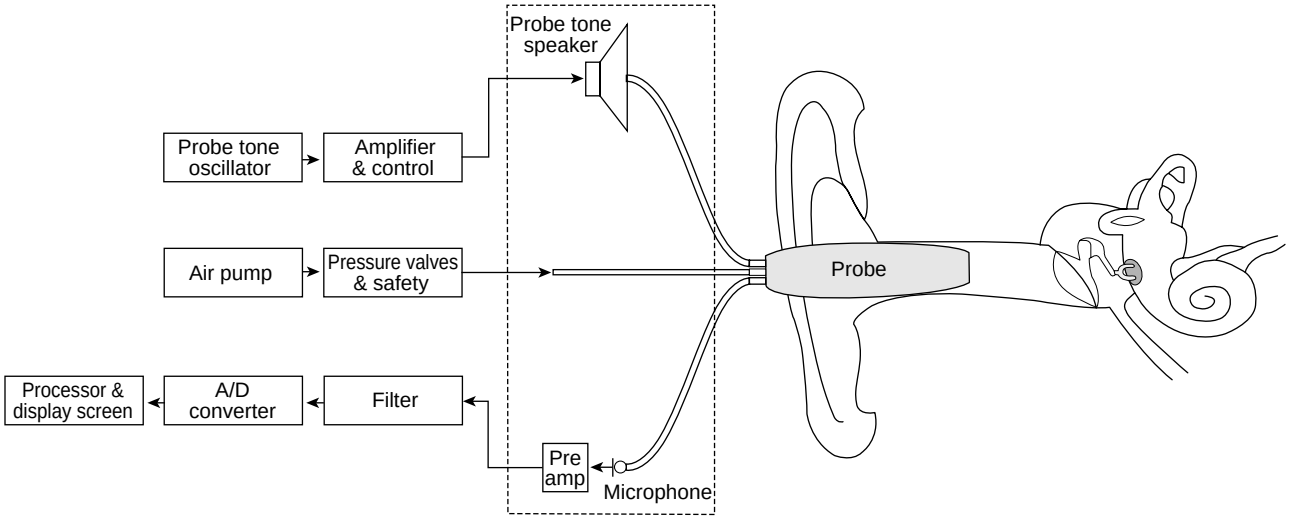


Figure 11. Block diagram of an acoustic immittance measuring device.

static acoustic immittance is obtained by adjusting the air pressure in the external ear canal so that a peak in the tympanogram exists. The magnitude of this peak, relative to the uncompensated immittance value, is clinically useful, because it can be compared with norms (e.g., 0.3 to 1.6 mmho) to determine the presence of middle ear pathology. It is important to note that at ear canal pressures of + 200 daPa or more, the sound pressure level (SPL) in the ear canal is directly related to the volume of air in the external ear canal, because the contribution of the middle ear system is insignificant at that pressure. A measure of the external ear canal volume is a valuable measure that can be used to detect tympanic membrane perforations otherwise difficult to detect visually. That is, a large ear canal volume (i.e., a value considerably greater than 1.5 mL) indicates a measurement of both the external ear canal and the middle ear as a result of a perforation in the tympanic membrane.

Dynamic Acoustic Immittance (Tympanometry)

The sound pressure of the probe-tone directed at the eardrum is maintained at a constant level and the volume velocity is measured by the instrumentation while positive and negative air pressure changes are induced in the external ear canal.

The procedure is called tympanometry and the resulting changes in immittance are recorded graphically as a tympanogram. A typical tracing is seen in Fig. 12.

Admittance is at its maximum when the pressures on both sides of the tympanic membrane are equal. Sound transmission decreases when pressure in the ear canal is greater or less than the pressure at which maximum admittance occurs. As a result, in a normal ear, the shape of the tympanogram has a characteristic peaked shape (see Fig. 13) with the peak of admittance occurring at an air pressure of 0 decapascals (daPa).

Tympanograms are sometimes classified according to shape (Fig. 13) (35).

The Type A tympanogram shown in Fig. 13, so-called because of it's resemblance to the letter "A", is seen in normal ears. When middle ear effusion is present, the fluid contributes to a decrease in admittance, regardless of the changes of pressure in the external ear canal. As a result, a characteristically flat or slightly rounded Type B tympanogram is typical. When the Eustachian tube malfunctions, the pressure in the middle ear can become negative relative to the air pressure in the external auditory canal. As energy flow through the ear is maximal when the pressure differential across the tympanic membrane is zero, tympanometry reveals maximum admittance when the pressure being

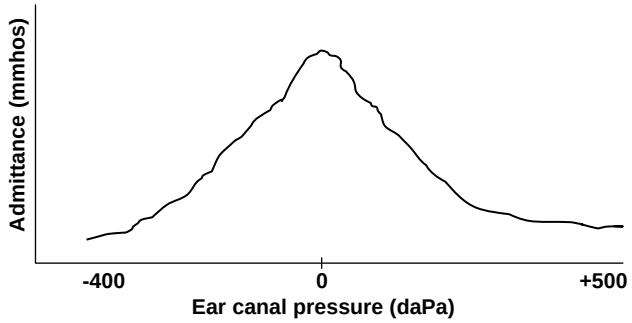


Figure 12. Tympanogram of a normal ear.

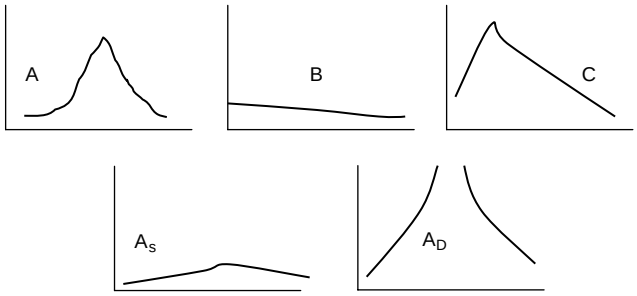


Figure 13. Tympanogram types (see text for descriptions).

varied in the external ear canal matches the negative pressure in the middle ear. At that pressure, the peak admittance will be normal but will occur at an abnormal negative pressure value. This tympanogram type is termed a Type C. It is important to note that variations exist of the type A tympanogram associated with specific pathophysiology affecting the middle ear. For example, if the middle ear is unusually stiffened by ear disease, the height of the peak may be reduced (Type A_s). Similarly, if middle ear pathology such as a break in the ossicular chain occurs, the energy flow may be enhanced, which is reflected in the Type A_d tympanogram depicted in Fig. 13.

Multifrequency Tympanometry

Under certain circumstances, particularly during middle ear testing of newborns and in certain stages of effusion in the middle ear, responses to the 226 Hz probe-tone typically used in tympanometry may fail to reveal immittance changes caused by disorders of the middle ear. In these circumstances, tympanometry with probe-tone frequencies above 226 Hz may be very useful in the detection of middle ear dysfunction. With probe-tone frequencies above 226 Hz, tympanometric shapes are more complex. More specifically, multifrequency tympanometric tracings normally progress through an expected sequence of shapes as probe frequency increases (36), and deviations from the expected progression are associated with certain pathologies.

Acoustic Reflex Measurement

In humans, a sufficiently intense sound causes a reflexive contraction of the middle ear muscles in both ears, acoustically stiffening the middle ear systems in each ear, called the acoustic reflex and is a useful tool in the audiometric test battery. When the reflex occurs, energy flow through both middle ears is reduced, and the resulting change in immittance can be detected in the probe ear by an immittance measuring device. Intense tones can be introduced to the probe ear (ipsilateral stimulation) or by earphone to the ear opposite the probe ear (contralateral stimulation).

One acoustic reflex measure is the minimum sound intensity necessary to elicit the reflex. The minimum sound pressure level necessary to elicit the reflex is called the acoustic reflex threshold. Acoustic reflex thresholds that are from 70 to 100 dBHL are generally considered to be in the normal range when pure tone stimuli are used. In general, the acoustic reflex thresholds in response to broadband noise stimuli tend to be lower than those for pure tones. Reduced or elevated thresholds, as well as unusual acoustic reflex patterns, are used by audiologists to localize the site of lesion and as one method of predicting auditory sensitivity.

OTOACOUSTIC EMISSIONS

When sound is introduced to the ear, the ear not only is stimulated by sound, it can also generate sounds that can be detected in the ear canal. The generated sounds, so-called otoacoustic emissions, have become the basis for the development of another tool that audiologists can use to assess the auditory system. In the following section, otoacoustic emissions will be described, and their relationship to conventional audiometry will be discussed.

Otoacoustic Emissions—Historical Perspective

Until relatively recently, the cochlea was viewed as a structure that converted mechanical energy from the middle ear into neural impulses that could be transmitted to and used by the auditory nervous system. This conceptual role of the cochlea was supported by the work on human cadavers of Georg von Bekesy during the early and middle 1900s, and summarized in 1960 (37). In Nobel Prize-winning research, von Bekesy developed theories to account for the auditory system's remarkable frequency sensitivity, and his views were widely accepted. However, a different view of the cochlea was proposed by one of von Bekesy's contemporaries, Thomas Gold, who suggested that processing in the cochlea includes an active process, a mechanical resonator (38). This view, although useful in explaining cochlear frequency selectivity, was not widely embraced at the time it was proposed.

In later years, evidence in support of Gold's idea of active processing in the cochlea accumulated. Particularly significant were direct observations of outer hair cell motility (39). In addition, observed differences in inner hair cell and outer hair cell innervation such as direct efferent innervation of outer but not inner hair cells (40) suggested functional differences in the two cell types. Most relevant to the present discussion were reports of the sounds that were recorded in the ear canal (41) and attributed to a mechanical process occurring in the cochlea, which are now known as otoacoustic emissions.

Otoacoustic Emissions—Description

Initially, otoacoustic emissions (OAEs) were thought to originate from a single mechanism, and emissions were classified on the basis of the stimulus conditions under which they were observed. For example, spontaneous otoacoustic emissions (SOAEs) are sounds that occur spontaneously without stimulation of the hearing mechanism. Two categories of otoacoustic emissions that are most widely used clinically by audiologists are (*1*) transient otoacoustic emissions (TOAEs), which are elicited by a brief stimulus such as an acoustic click or a tone burst, and (*2*) distortion product otoacoustic emissions (DPOAEs), which are elicited by two tones (called primaries) that are similar, but not identical, in frequency. A third category of otoacoustic emissions that may prove helpful to audiologists in the future is the stimulus frequency otoacoustic emission (SFOAE), which is elicited with a pure tone. Currently, SFOAEs are used by researchers studying cochlear function, but they are not used widely in clinical settings.

Recent research indicates that, contrary to initial thinking, otoacoustic emissions are generated by at least two mechanisms, and a separate classification system has been proposed to reflect improved understanding of the physical basis of the emissions. Specifically, it is believed that the

mechanisms that give rise to evoked otoacoustic emissions include (*1*) a nonlinear distortion source mechanism and (*2*) a reflection source that involves energy reflected from irregularities within the cochlea such as variations in the number of outer hair cell motor proteins or spatial variations in the number and geometry of hair cell distribution (42). Emissions currently recorded in the ear canal for clinical purposes are thought to be mixtures of sounds generated by these two mechanisms.

Instrumentation

Improved understanding of the mechanisms that generate otoacoustic emissions may lead to new instrumentation that can "unmix" evoked emissions. Currently, commercially available clinical equipment records "mixed" emissions and includes a probe placed in the external ear canal that both delivers stimuli (i.e., pairs of primary tonal stimuli across a broad range of frequencies, clicks or tone-bursts) and records resulting acoustic signals in the ear canal. The microphone in the probe equipment is used in (*1*) the verification of probe fit, (*2*) monitoring probe status (e.g., for cerumen occlusion), (*3*) measuring noise levels, (*4*) verifying stimulus characteristics, and (*5*) detecting emissions. Otoacoustic measurement recording entails use of probe tips of various sizes to seal the probe in the external ear canal and hardware/software that control stimulus parameters and protocols for stimulus presentation. The computer equipment performs averaging of responses time-locked to stimulus presentation, noise measurement, artifact rejection, data storage, and so on, and can provide stored normative data and generate printable reports. An example of a typical DPOAE data display is shown in Fig. 14.

It is important to note that outer or middle ear pathology can interfere with transmission of emissions from the cochlea to the ear canal, and thus the external ear canal and middle ear status are important factors in data interpretation. Also, although otoacoustic emissions ordinarily are not difficult to record and interpret, uncooperative patient behavior and high noise levels can hamper or even preclude measurement of accurate responses.

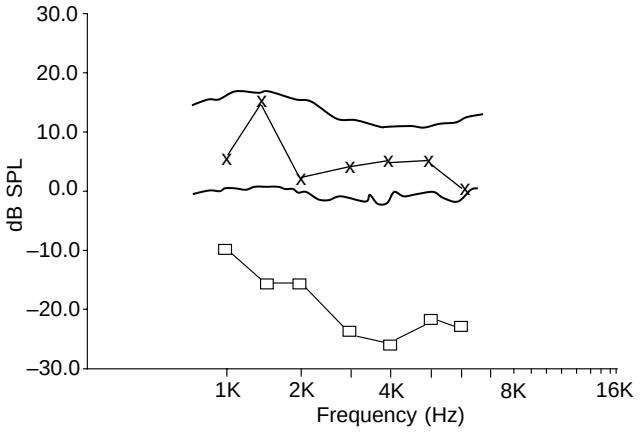


Figure 14. DPAOE responses from 1000 to 6,000 Hz from the left ear of a normal listener. X = DPOAE response amplitudes; squares = physiologic noise floor; bold curves = 95% confidence limits for normal ears.

Otoacoustic Emissions—Clinical Applications

Measurement of otoacoustic emissions is used routinely as a test battery component of audiometric evaluations in children and adults, and it is particularly useful for monitoring cochlear function (e.g., in cases of noise exposure and during exposure to ototoxic medication) as well as differentiating cochlear from neural pathology. Currently, otoacoustic emission evaluation is also useful, either alone or in combination with evoked potential recording, in newborn hearing screening. In addition, OAE assessment is sometimes used in preschool and school-age hearing screening, as well as with patients who may be unwilling to cooperate during audiometry.

Otoacoustic Emissions and Audiometric Threshold Prediction/Estimation

Audiologists do not use otoacoustic emissions as a measure of "hearing" because OAEs constitute an index of cell activity in the inner ear, not "hearing." Research suggests, however, that otoacoustic emissions may become an important indicator for predicting/estimating auditory thresholds when conventional audiometry cannot be conducted (42).

Many sources of variability exist that affect DPOAE use in audiometric threshold prediction/estimation, including variability with respect to etiology of the hearing loss, age, gender, and uncertainty regarding locations of DPOAE generation and their relationship to audiometric test frequencies. Individual DPOAE amplitude variation, intra- and inter-subject variations occurring at different frequencies and at different stimulus levels, and the mixing of emissions from at least two different regions of the cochlea (as described above) can reduce frequency selectivity and specificity in DPOAE measurement.

It has been suggested that developing methods to "unmix" the emissions associated with different generators (e.g., through the use of suppressor tones to reduce or eliminate one source component or with the use of Fourier analysis to analyze the emissions) may reduce variability and improve specificity in threshold prediction/estimation and determination of etiology (41). It is likely that future commercial otoacoustic measurement instruments will enable the user to differentiate distortion source emissions from reflection source emissions and that this improvement will lead to more widespread use of otoacoustic emissions in audiometric threshold estimation and prediction.

BIBLIOGRAPHY

1. American Academy of Audiology. Scope of practice. Audiol Today 2004;15(3):44–45.
2. American National Standards Institute. Methods of manual pure tone threshold audiometry. (ANSI S3.21-2004). New York: ANSI; 2004.
3. Bergman M. On the origins of audiology: American wartime military audiology. Audiol Today Monogr 2002;1:1–28.

4. Seashore CE. An Audiometer. University of Iowa Studies in Psychology (No. 2). Iowa City, NA: University of Iowa Press; 1899.
5. Bunch CC. Clinical Audiometry. St. Louis, MO: C. V. Mosby; 1943.
6. Fowler EP, Wegel RL. Presentation of a new instrument for determining the amount and character of auditory sensation. Trans Am Otol Soc 1922;16:105–123.
7. Bunch CC. The development of the audiometer. Laryngoscope 1941;51:1100–1118.
8. Carhart R. Clinical application of bone conduction audiometry. Archi Otolaryngol 1950;51:798–808.
9. American National Standards Institute. Specification of audiometers. (ANSI-S3.6-1996). New York: ANSI; 1996.
10. American National Standards Institute. Maximum permissible ambient noise levels for audiometric test rooms. (ANSI-S3.1-1999 [R2003]). New York: ANSI; 1999.
11. Breakwell GM, Hammond S, Fife-Shaw C. Research Methods in Psychology. 2nd ed. London: Sage; 2000. p 200–205.
12. von Bekesy G. A new audiometer. Acta Oto-laryngologica Stockholm 1947;35:411–422.
13. Jerger J. Bekesy audiometry in analysis of auditory disorders. J Speech Hear Res 1960;3:275–287.
14. Jerger J, Herer G. Unexpected dividend in Bekesy audiometry. J Speech Hear Disord 1961;26:390–391.
15. Jerger S, Jerger J. Diagnostic value of Bekesy comfort loudness tracings. Arch Otolaryngol 1974;99:351–360.
16. Carhart R, Jerger J. Preferred method for clinical determination of pure-tone thresholds. J Speech Hear Disord 1959;24: 330–345.
17. Feldmann H. A history of audiology: A comprehensive report and bibliography from the earliest beginnings to the present. Translations Beltone Institute Hear Res 1960;22:1–111. [Translated by Tonndorf J. from Die Geschichtliche Entwicklung der Horprufungsmethoden, Kuze Darstellung and Bibliographie von der Anfrongen bis zur Gegenwart. In: Leifher H, Mittermaiser R, Theissing G, editors. Zwanglose Abhandunger aus dem Gebeit der Hals-Nasen-Ohren-Heilkunde. Stuttgart: Georg Thieme Verlag; 1960.]
18. Hudgins CV, Hawkins, JE Jr, Karlin JE, Stevens SS. The development of recorded auditory tests for measuring hearing loss for speech. Laryngoscope 1947;57:57–89.
19. Olsen WO, Matkin ND. Speech audiometry. In: Rintelmann WF, editor. Hearing Assessment. 2nd ed. Austin, TX: Pro-Ed; 1991. 39–140.
20. Fletcher H. Speech and Hearing. Princeton NJ: Von Nostrand; 1929.
21. Fletcher H. A method of calculating hearing loss for speech from an audiogram. Acta Otolaryngologica 1950; (Suppl 90): 26–37.
22. Carhart R. Observations on the relations between thresholds for pure tones and for speech. J Speech Hear Disord 1971; 36:476–483.
23. Beattie RC, Edgerton BJ, Svihovec DV. An investigation of Auditec of St. Louis recordings of Central Institute for the Deaf spondees. J the Am Audiol Soc 1975;1:97–101.
24. Cambron NK, Wilson RH, Shanks JE. Sondaic word detection and recognition functions for female and male speakers. Ear Hear 1991;12:64–70.
25. Loomis A, Harvey E, Hobart G. Disturbances of patterns in sleep. J Neurophysiol 1938;1:413–430.
26. Clark WA Jr, Goldstein MH Jr, Brown RM, Molnar CE, O'Brien DF, Zieman HE. The average response computer (ARC): A digital device for computing agerages and amplitudes and time histograms of physiological responses. Trans of IRE 1961;8:46–51.
27. Jewett DL, Romano MN, Williston JS. Human auditory evoked potentials: Possible brainstem components detected on the scalp. Science 1970;167:1517–1518.
28. Jewett DL, Williston JS. Auditory evoked far fields averaged from the scalp of humans. Brain 1971;94:681–696.
29. Stapells DR, Oates P. Estimation of the pure-tone audiogram by the auditory brainstem response: A review. Audiol Neurootol 1997;3(5):257–280.
30. Rance G, Briggs RJS. Assessment of hearing in infants with moderate to profound impairment: The Melbourne experience with auditory stead-state evoked potential testing. Ann Otol Rhinol Laryngol 2002;111(5) (Part 2, Suppl. 189):22–28.
31. Swanepoel D, Roode R. Auditory steady-state responses for children with severe to profound hearing loss. Arch of Otolaryngol Head Neck Surg 2004;130(5):531–535.
32. Margolis RH, Hunter LH. Acoustic Immittance Measurements. In: Roeser RJ, Valente M, Hosford-Dunn H, editors. Audiology Diagnosis. New York: Thieme; 2002. pp 381–423.
33. Terkildsen K, Nielsen S. An electroacoustic impedance measuring bridge for clinical use. Arch Otolaryngol 1960;72:339–346.
34. American National Standards Institute. National Standard Specifications for instruments to measure aural acoustic impedance and admittance. (ANSI S3.39-1987). New York: ANSI; 1987.
35. Jerger J. Clinical experience with impedance audiometry. Arch Otolaryngol 1970;92:311–324.
36. Vanhuyse VJ, Creten WL, Van Camp KJ. On the W-notching of tympanograms. Scand Audiol 1975;4:45–50.
37. von Bekesy G. Experiments in Hearing. New York: McGraw Hill; 1960.
38. Gold T. Hearing. II. The physical basis of the acton of the cochlea. Proc Roy Soc Brit 1948;135:492–498.
39. Brownell WE, Bader CR, Bertrand D, Ribaupierre Y. Evoked mechanical responses of isolated cochlear outer hair cells. Science 1985;227:194–196.
40. Smith CA. Innervation pattern of the cochlea. Ann Otol Rhinol Otolaryngol 1961;70:504–527.
41. Kemp DT. Stimulated acoustic emissions from within the human auditory system. J Acoust Soc Am 1978;64:1386–1391.
42. Shera CA. Mechanisms of mammalian otoacoustic emission and their implications for the clinical utility of otoacoustic emissions. Ear Hear 2004;25:86–97.

See also COCHLEAR PROSTHESES; COMMUNICATION DEVICES.

AUDITORY IMPLANTS.

See COCHLEAR PROSTHESES.

AUGMENTATIVE COMMUNICATION SYSTEM.

See COMMUNICATION DEVICES.

B

BACTERIAL DETECTION SYSTEMS. See MICROBIAL DETECTION SYSTEMS.

BALLOON PUMP. See INTRAAORTIC BALLOON PUMP.

BANKED BLOOD. See BLOOD COLLECTION AND PROCESSING.

BAROTRAUMA. See HYPERBARIC MEDICINE.

BARRIER CONTRACEPTIVE DEVICES. See CONTRACEPTIVE DEVICES.

BIOCERAMICS. See BIOMATERIALS: BIOCERAMICS.

BIOCOMPATIBILITY OF MATERIALS

ZHOU XIANG
MYRON SPECTOR
Brigham and Women's Hospital
Boston, Massachusetts

INTRODUCTION

Most surgical specialties have been revolutionized by the introduction of implantable devices. These advances have been founded, in large part, on biomaterials science and engineering. One of the critical determinants of the performance of the device relates to its compatibility with the structure and function of the tissue or organ (a structure comprised of two or more tissues) in which it is implanted. Moreover, the tissue response to the implant should not impede the required function of the device. This article will deal with the response to biomaterials implanted into solid tissues. Biocompatibility issues related to blood-contacting applications will be outside the scope of this discussion.

Implantation of the device requires the production of a surgical wound, and in this respect the tissue response to the implant may be looked upon as the modification of the wound healing response by the very presence of the implant. In vascularized tissues, the creation of a surgical wound elicits an inflammatory process that can be considered part of the natural course of healing. The end result of the healing process is tissue similar to that naturally occurring at the site (the process of "regeneration") or scar tissue (the process of "repair"), which in many tissues and organs comprises fibrous tissue. Infectious organisms (viz., bacteria) when present serve as a persistent injurious agent that prolongs and can further incite the inflammatory process, not only jeopardizing the performance of the implant, but threatening the life of the individual. The principles of biocompatibility including the mechanisms underlying inflammatory and infectious processes apply regardless of the type of material of fabrication of the implant. There are, however, features of the biomaterials that can affect certain aspects of these processes. It can therefore be instructive to also consider issues of biocompatibility in the context of the various classes of materials: metals, ceramics, and polymers.

CLASSES OF MATERIALS

The term "biomaterials" generally refers to synthetic materials and treated natural materials that are employed for the fabrication of implantable devices that are to replace or augment tissue or organ function. An understanding of the chemical makeup of biomaterials can provide insights into their biocompatibility for selected applications. Generally, biomaterials may be considered "inert" or "reactive" with the biological milieu. In the latter case, the reactivity could relate to the release of moieties or the adsorption of biological molecules. Inert materials may also release small amounts of ions and molecules or nonspecifically bind biomolecules. The feature that distinguishes inert from reactive biomaterials is the degree to which the interactions of the implant with the biological environment affects the tissue response and device performance. Those materials designed to effect specific tissue responses through their reactivity may also be referred to as "bioactive".

This article provides a brief description of materials used for the fabrication of implantable devices relative to issues related to biocompatibility. A more comprehensive discussion of biomaterials can be found elsewhere in this encyclopedia.

Metals

In metals, closely packed arrays of positively charged atoms are held together in a loosely associated "cloud" of free electrons. The essential features of the metallic bond are that it is nondirectional and the electrons are freely mobile. The metals most often used for the fabrication of implantable devices are stainless steel, cobalt–chromium alloys, and titanium and titanium alloy. The specific members of these families used as biomaterials are usually identified by a designation provided by the American Society for Testing and Materials (ASTM). Metallic materials have certain properties that make them ideal for load-bearing applications; in particular, they can maintain very high strength under the aggressive aqueous environment in the body.

The biocompatibility of the implantable metallic materials is related to their corrosion resistance, in that they can generally be considered as inert. While they release detectable levels of metal ions (1,2), these ions have not yet been found to significantly affect tissue or organ function or cause pathological changes. One conundrum related to biomaterials is that while the ions released from certain metallic devices are known to be carcinogenic when administered to certain animals models (3,4) and when encountered by humans in certain circumstances (1), there have not yet been definitive studies relating the incidence of

various cancers to the ions released by metallic implants (see later section).

The following sections provide a brief description of three metal systems most frequently used for the fabrication of implantable devices.

Stainless Steels. The stainless steels, like all steels, are iron-based alloys. Chromium is added to improve the corrosion resistance through the formation of a chromium-oxide surface layer; at least 17% chromium is required for the term "stainless" to be used. Carbon and nickel are employed as alloying elements to increase strength. The most common type of stainless steel used for implants is 316L (American Iron and Steel Institute designation; ASTM F-138), containing 17–19% chromium, 13–15.5% nickel, and <0.03% carbon. Despite their very good corrosion resistance, stainless steels are subject to several types of corrosion processes, including crevice, pitting, intergranular, and stress corrosion. These processes can profoundly degrade the mechanical strength of the alloy and can lead to the elevated release of metallic ions into the surrounding tissue with undesirable biological consequences.

Alloying with chromium generates a protective, self-regenerating oxide film that resists perforation, has a high degree of electrical resistivity, and, thus, provides a major protection against corrosion; formation of the chromium oxide "passivation" layer is facilitated by immersion of the alloy in a strong nitric acid solution. The nickel imparts more corrosion resistance and ease of fabrication. The molybdenum addition provides resistance to pitting corrosion. Other alloying elements facilitate manufacturing processes.

Cobalt–Chromium Alloys. Surgical cobalt–chromium alloy is a cobalt-based system with chromium added for increased corrosion resistance. Its composition contains 27–30% chromium and 5–7% molybdenum. Tungsten is added to the wrought alloy to enhance ductility. As with stainless steels, the chromium content of this alloy generates a highly resistive passive film that contributes substantially to corrosion resistance. The Co–Cr Mo (F-75) alloy has superior corrosion resistance to the F-138 stainless steel, particularly in crevice corrosion. There is an extensive, decades-long history of biocompatibility in human implantation. The Co–Cr Mo devices are currently produced by a process referred to as hot isostatic pressing that results in parts with more favorable strength characteristics than results from casting processes. The Co–Cr Tn–Ni Alloy (ASTM F90) is very different from the F-75 alloy with which it is often confused. Parts can be fabricated from this alloy by hot forging and cold drawing; it is not used in the cast form. In clinical practice, F90 is used to make wire and internal fixation devices (e.g., plates, intramedullary rods, and screws).

Unalloyed Titanium (ASTM F-67) and Titanium Alloy (ASTM F-136). Titanium and its alloy with 6% aluminum and 4% vanadium, Ti-6A1-4V, are used for their excellent corrosion resistance and their modulus of elasticity that is approximately one-half that of stainless steel and cobalt–chromium alloys. This lower modulus results in devices with lower stiffness that may be advantageous in certain applications such as implants in bone as they will result in less stress-shielding of bone. The alloy of titanium has much better material properties than the pure titanium. Problems with titanium are its severe notch sensitivity and poor wear resistance.

Titanium and its alloys are of particular interest for biomedical applications due to their outstanding biocompatibility. In general, their corrosion resistance significantly exceeds that of the stainless steels and the cobalt–chromium alloys. In saline solutions near neutral pH, the corrosion rate is extremely small, and there is no evidence of pitting, intergranular, or crevice corrosion. Unalloyed titanium is used less frequently than the alloy for the fabrication of implants. It is, however, available in various configurations, such as plain wire for manufacturing purposes. In addition, it is used to produce porous coatings for certain designs of total joint replacement prostheses.

ASTM F-136 specifies a titanium alloy with a content of 5.5–6.5% Al, 3.5–4.5% V, 0.25% Fe (maximum), 0.05% N (maximum) and 0.08% C (maximum), 0.0125% H (maximum), and other 0.1% (maximum 0.4% total). Developed by the aircraft industry as one of a number of high strength titanium alloys, this particular formulation has a yield strength reaching 1110 MPa. The ASTM F-136 specification limits the oxygen to an especially low level of 0.13% maximum. This is also known in the industry as the ELI (extra low interstitial) grade. Limiting the level of oxygen improves the mechanical properties of the material, particularly increasing its fatigue life. One interesting feature of titanium and its alloys is the low modulus of elasticity of 100 GPa as compared to 200 GPa for the cobalt–chromium alloys. This feature leads to their use in plates for internal fixation of fractures. Some have found that the lower stiffness of these plates may decrease the severity of bone stress-shielding that results in osteopenia under these devices.

One of the weaknesses of titanium is its poor wear resistance. It appears that this problem relates to the mechanical stability of the passive film covering the surface of the alloy. On a carefully polished surface, the film is highly passive but mechanically weak. The poor wear resistance of titanium can result in the release of particulate wear debris if the material is not judiciously employed in the fabrication of implants (5). While the metal in bulk form may be biocompatible adverse cell and tissue responses may be elicited by titanium particles (5–8).

Permanent and Absorbable Synthetic and Natural Polymers

Polymers consist of long chains of covalently bonded molecules characterized by the repeated appearance of a monomeric molecular unit. They can be produced de novo by the polymerization of synthetic monomers or prepared from natural polymers isolated from tissues. Most synthetic and natural polymers have a carbon backbone. Bonding among polymer chains results from much weaker secondary forces—hydrogen bonds or van der Waal's forces. Covalent bonding among chains, referred to as cross-linking, can be produced in certain polymer systems. Physical

entanglements of the long polymer chains, the degree of crystallinity, and chemical cross-linking among chains play important roles in determining polymer properties. The molecular bonding of the backbone of the polymer can be designed to undergo hydrolysis or enzymatic breakdown thus allowing for the synthesis of absorbable, as well as permanent, devices.

Polymeric materials are generally employed for the fabrication of implants for soft tissue applications that require a greater degree of compliance than can be achieved with metals. However, they have also been shown to be of value as implants in bone for indications that would also benefit from their lower modulus of elasticity, and the ability of some to be polymerized *in vivo* so as to adapt to defects of complex shape. For some indications, the radiolucency of polymeric materials may be an important benefit. Because of the limited strength and wear resistance of polymers, care must be given to the load-bearing requirements of the applications in which they are used.

The following sections provide a summary of a few of the most frequently used polymers.

Polymethyl Methacrylate (PMMA). Polymethyl methacrylate (9) is used in a self-curing form as a filling material for defects in bone and as a grouting agent for joint replacement prostheses. It can be shaped *in vivo* while in a dough stage prior to complete polymerization and thus makes a custom implant for each use. Its purpose is the redistribution of stress in a more even pattern to the surrounding bone. Often referred to as "bone cement", when PMMA is employed for joint arthroplasty it acts as a grout to support the prosthesis rather than a glue; it has minimal adhesive properties. The time-dependent properties of PMMA during curing require an understanding of its handling characteristics. Immediately after mixing, the low viscosity permits interdigitation with cancellous bone. Viscosity rises quickly once the chemical setting reaction begins requiring that the prosthesis be accurately positioned and stationary to achieve maximum fixation.

The chemical toxicity of the methyl methacrylate monomer and the heat generated during the polymerization exothermic reaction need be considered when using PMMA. While the toxicity of the monomer has not prevented the material from being employed successfully for a wide array of applications, there are efforts to reduce the monomer content (10) and to employ alternative activating agents that would be less toxic (11).

Its brittle nature after curing and low fatigue strength make PMMA vulnerable to fracture under high mechanical loading and production of wear debris in situations when other harder materials rub against it. The cellular response to PMMA particles can promote an inflammatory response (12,13) that can result in osteolysis. This process has been referred to as "cement disease" (14). This underscores the importance of reducing the circumstances that can result in the production of PMMA debris.

Silicone. Silicones (15) are polymers having a backbone comprising alternating silicon and oxygen atoms with organic side groups bonded to the silicon through covalent bonding with the carbon atom. One form of silicone commonly used for the fabrication of implants is polydimethylsiloxane (PDMS). In PDMS, methyl (CH_3) side groups are covalently bonded to the silicon atom and it can be used in three forms: (a) a fluid comprising linear polymers of varying molecular weight (i.e., chain length); (b) a cross-linked network referred to as a gel; and (c) a solid elastomer comprising a highly cross-linked gel filled with small particles of silica. In considering the performance of silicone implants, the role of each form of PDMS in the device need be considered. Attributing specific biological responses to individual components of a silicone device is complicated by the many molecular forms of silicone that the implant comprises.

The PDMS elastomers contain a noncystalline silica particles 7–22 nm in diameter that has been surface-treated to facilitate chemical bonding of the particle to the PDMS gel. Addition of the silica particle to a highly cross-linked PDMS gel is done to modify the mechanical properties of the elastomer. One of the challenges in investigating tissue responses that may have been elicited by the release of the silica particles is their small size that requires transmission electron microscopy (TEM) for analysis.

Polyethylene. Ultrahigh molecular weight polyethylene (UHMWPE) (16,17) has a very low frictional coefficient against metal and ceramics, and is therefore used as a bearing surface for joint replacement prostheses. Moreover, the wear resistance of UHMWPE is higher than other polymers investigated for this application. Low strength and creep, however, present potential problems.

The term polyethylene refers to plastics formed from polymerization of ethylene gas. The possibilities for structural variation on molecules formed by this simple repeating unit for different molecular weight, crystallinity, branching, cross-linking, and so on, are so numerous and dramatic with such a wide range of attainable properties that the term polyethylene truly refers to a subclass of materials. The earliest type of polyethylene was made by reacting ethylene at high (20,000–30,000 lb/in.^{2}) pressure and temperatures of 200–400 °C with oxygen as catalyst. Such material is referred to as conventional or low density polyethylene. Much polyethylene is produced now by newer, low pressure techniques using aluminum–titanium (Ziegler) catalysts. This is called linear polyethylene due to the linearity of its molecules, in contrast to the branched molecules produced by high pressure processes. The linear polymers can be used to make high density polyethylene by means of the higher degree of crystallinity attained with the regularly shaped molecules. Typically, there is no great difference in molecular weight between the low density and high density varieties, (e.g., 100,000–500,000). However, if the low pressure process is used to make extremely long molecules, (i.e., UHMWPE), the result is different and quite remarkable. This material, with a molecular weight between 1 and 10 million, is less crystalline and less dense than high density polyethylene and has exceptional mechanical properties. The material is used in very demanding applications and is by far the most successful polymer used in total joint replacements. It far outperforms the various acrylics, fluorocarbons,

polyacetals, polyamides, and polyesters that were tried for such purposes. In recent years, cross-linking methods including chemical agents and ionizing radiation have been implemented in an attempt to further improve the wear performance of UHMWPE.

The principal biocompatibility issue with polyethylene relates to the inflammatory response provoked by particles of the polymer, as can be generated by the wear of total joint replacement prostheses prostheses (18,19).

Absorbable Polymers. Absorbable polymers have been used in the fabrication of surgical implants for decades in the form of absorbable sutures. More recently, this class of materials has been investigated for the application of resorbable devices including fracture fixation implants and scaffolds for tissue engineering. The principal issues associated with the implementation of absorbable polymers as implants include: mechanical properties (i.e., strength), degradation rate, and biological response to the degradation products.

Synthetic Absorbable Polymers. One of the classes of polymers used frequently for the fabrication of absorbable implants is the alpha-hydroxy acids including L-lactic acid, glycolic acid, and dioxanone. These molecules normally are used in their polymeric forms: poly L-lactic acid (PLLA), polyglycolic acid (PGA), and polydioxanone. Copolymers of lactic and glycolic acids are also frequently employed.

This particular class of polyester undergoes breakdown as a result of the hydrolytic scission of the ester bond. The access of water to this bond in PGA is much greater resulting in a more rapid degradation rate compared to that which occurs with PLLA, which has a bulkier CH_3 side group instead of the H atom in PGA. The copolymer of polylactic and polyglycolic acid can be designed to have an intermediate degradation rates. While the majority of the breakdown of these polymers is due to hydrolytic scission there is some lesser extent of nonspecific enzymatic action.

Several factors affect the rate of breakdown of these polymers: the relative amount of monomers comprising the copolymers, the degree of crystallinity, and the surface area. These polymers are normally broken down to natural body components excreted in the urine or exhaled. The process of degradation involves the gradual decrease in the average molecular weight of the polymer as hydrolysis proceeds. At some point, the molecular weight decreases to the extent that the polymer becomes soluble in the aqueous environment and there is a bolus release of the molecules. Depending on the mass of the implanted device, the concentration of the molecules may elicit an inflammatory response (20).

Natural Absorbable Polymers. Myriad devices are fabricated from collagen (21), the principal structural protein of the body. The collagen molecule comprises three tightly coiled helical polypeptide chains. *In vivo* the collagen molecule, tropocollagen, is assembled to form fibrils that in turn assume various orientations and configurations to form the architecture of various tissues. The wide array of properties of tissues comprising collagen, from dermis to musculoskeletal tissues including articular cartilage, meniscus, and ligament, is due to differences in the chemistry, density, and orientation of the fibrils formed from the collagen molecule.

Collagen is soluble in specific solutions in which the chains can become disentangled to produce gelatin. It can be isolated from tissue and purified through the use of several agents: acids, alkalis, enzymes, and salt. Treatment in acid results in the elimination of acidic proteins and glycosaminoglycans that result in the dissociation of the collagen fibrils. A similar effect can be achieved using alkaline extraction with the removal of basic proteins. Proteolytic enzymes that cleave the telopeptides, that serve as natural cross-linking agents for collagen, allow for the dissolution of collagen molecules and aggregates in aqueous solutions. Salt extraction leads to the removal of newly synthesized collagen molecules and certain noncollagenous molecules thus facilitating the disaggregation of collagen fibrils.

That collagen is soluble in acidic medium facilitates its extraction from tissues and reprocessing into biomaterials. Several factors are critical determinants of the properties of reconstituted collagen biomaterials. The degree to which denaturation or degradation of the collagen structures isolated from tissue occurs will affect the mechanical properties. These properties will also be affected by the degree to which the material is subsequently cross-linked.

An important biological property related to the molecular structure of collagen is the collagen-induced blood platelet aggregation. The quarternary structure of collagen resulting from the periodic aggregation of the collagen molecules has been well documented. Methods for isolating and purifying collagen fibrils, that result in the preservation or destruction of this quarternary structure are employed to produce either hemostatic or thromboresistant biomaterials. Another factor relates to the removal of soluble components that might serve as antigens. The immunogenicity can be reduced, to clinically nonconsequential levels, by chemically modifying the antigen molecules.

A wide variety of methods have been employed for the fabrication of collagen sutures, fleeces for hemostasis, and sponge-like materials for scaffolds for tissue engineering.

Ceramics

Ceramics are typically three-dimensional (3D) arrays of positively charged metal ions and negatively charged nonmetal ions, often oxygen. The ionic bond localizes all the available electrons in the formation of a bond. Network organization ranges from highly organized, crystalline, 3D arrays to amorphous, random arrangements in glassy materials.

Ceramics, for reasons described above, may be the most chemically inert implant materials currently in use. However, their relatively low tensile strength, high modulus, and brittleness limit the applications in which they may be used. Current techniques allowing the formation of ceramic coatings on metallic substrates have revitalized interest in ceramics for hard tissue applications.

Aluminum Oxide. Aluminum oxide has been found of value for the articulating components of total joint arthroplasties because of its high wear resistance and its low coefficient of friction when prepared in congruent, polished geometries. The brittle nature of alumina remains a detriment.

Calcium Phosphates/Hydroxyapatite. Calcium-based ceramics, closely related to the naturally occurring hydroxyapatite in bone, have generated a large amount of interest in recent years. The ability to bond directly to bone as well as their osteoconductive capability promise to enhance biological fixation of implant devices. Hydroxyapatite is only slightly resorbable and is used in both dense and porous forms as a permanent implant. Tricalcium phosphate is bioabsorbable to varying degrees, depending on formulation and structure. There are currently a wide array of calcium phosphate materials undergoing investigation as bone graft substitute materials (see the Section on Calcium Phosphate Materials as Matrices for Bone Regeneration: Bone Graft Substitute Materials).

Composite Materials

Composite materials are combinations of two or more materials, and usually more than one material class (i.e., metals polymers, ceramics). They are used to achieve a combination of mechanical properties for specific applications. Composite technology, much of it developed for the aerospace industry, is beginning to make it way into biomedical materials. Carbon fiber reinforced polymers are being investigated as substitutes for metals. The advantage is that devices with comparable strength, but with significantly lower stiffness can be produced. Moreover, these types of composite devices are radiolucent.

BIOLOGICAL RESPONSE TO BIOMATERIALS

The biological processes comprising the tissue response are affected by implant-related factors including (22):

1. The "dead space" created by the presence of the implant.
2. Soluble agents released by the implant (e.g., metal ions or polymer fragments).
3. Insoluble particulate material released from the implant (e.g., wear debris).
4. Chemical interactions of biological molecules with the implant surface.
5. Alterations in the strain distribution in tissue caused by the mismatch in modulus of elasticity between the implant and surrounding tissue, and the movement of the implant relative to adjacent tissue as a result of the absence of mechanical continuity.

Study of the tissue response to implants requires methodology capable of measurements at the molecular, cellular, and tissue levels. Moreover, time is an important variable owing to the criticality of the temporal relationship between the molecular and cellular protagonists of the biological reactions, and because implant-related factors act with different time constants on the biological responses. The dynamic nature of implant-tissue interactions requires that the final assessment of tissue compatibility be qualified by the time frame in which it has been evaluated.

The tissue response to an implant is the cumulative physiological effect of (*1*) modulation of the acute wound healing response to the surgical trauma of implantation and the presence of the implant, (*2*) the subsequent chronic inflammatory reaction associated with the presence of the device, and (*3*) remodeling of surrounding tissue as it adapts to the presence of the implant (23). Moreover, the healing and stress-induced adaptive remodeling responses of different tissues vary considerably. In this regard, the response of various tissues to the same implant can vary greatly.

In considering the biological response that might be elicited by an implant, the healing–remodeling characteristics of the four basic types of tissue: connective tissue, muscle, epithelia, and nerve, should be recalled. The characteristics of the parenchymal cells in each type of tissue can provide a basis for understanding the tissue response to an implant. The following characteristics of an implant site are determinants of the biological response:

1. Vascularity.
2. The nature of the parenchymal cell with respect to its capability for mitosis and migration, because these processes determine the regenerative capability of the tissue.
3. The presence of regulatory cells such as macrophages/histiocytes.
4. The effect of mechanical strain, associated with deformation of the extracellular matrix, on the behavior of the parenchymal cell.

Surgical wounds in avascular tissue (e.g., the cornea and inner third of the meniscus) will not heal because of the limited potential for the proliferation and migration of surrounding parenchymal cells and the absence of a fibrin clot in the wound site into which the cells can migrate. Gaps between an implant and surrounding avascular tissue can remain indefinitely. Implant sites in vascular tissues in which the parenchymal cell does not have the capability for mitosis heal by the formation of scar in the gap between the implant and surrounding tissue. Moreover, adjacent cells that have died as a result of the implant surgery will be replaced by fibroblasts and scar tissue.

Normal Local Tissue Response

Wound Healing. Implantation of a medical device initiates a sequence of cellular and biochemical processes that lead to "healing by second intention" (i.e., healing by the formation of granulation tissue within a defect; as opposed to the healing of an incision, i.e., healing by first intention). The first phase of healing in vascularized tissues is inflammation, which is followed by a reparative phase, the replacement of the dead or damaged cells by healthy cells. The pathway that the reparative process takes depends

on the regenerative capability of the cells comprising the injured tissue (i.e., the tissue or organ into which the implant has been placed). Cells can be distinguished as labile, stable, or permanent based on their capacity to regenerate. Labile cells continue to proliferate throughout life, replacing cells that are continually being destroyed. Epithelia and blood cells are examples of labile cells. Cells of splenic, lymphoid, and hematopoietic tissues are also labile cells. Stable cells retain the capacity for proliferation, although they do not normally replicate. These cells can undergo rapid division in response to a variety of stimuli and are capable of reconstitution of the tissue of origin. Stable cells include the parenchymal cells of all of the glandular organs of the body (e.g., liver, kidney, and pancreas), mesenchymal derivatives such as fibroblasts, smooth muscle cells, osteoblasts and chondrocytes, and vascular endothelial cells. Permanent cells are those which cannot reproduce themselves after birth. Examples are nerve cells.

Tissues comprised of labile and stable cells have the capability for regeneration after surgical trauma. The injured tissue is replaced by parenchymal cells of the same type, often leaving no residual trace of injury. However, tissues comprised of permanent cells are repaired by the production of fibrocollagenous scar. Despite the capability of many tissues to undergo regeneration, destruction of the tissue stroma, remaining after injury or constructed during the healing process, will lead to formation of scar. The biological response to materials thereby depends on the influence of the material on the inflammatory and reparative stages of wound healing. Does the material yield leachables or corrosion products that interfere with the resolution of inflammation initiated by the surgical trauma? Does the presence of the material interfere with the stroma required for the regeneration of tissue at the implant site? These are the types of questions that need to be addressed when assessing the "biocompatibility" of materials.

A number of systemic and local factors influence the inflammatory-reparative response. Systemic influences include age, nutrition, hematologic derangements, metabolic derangements, hormones, and steroids. While there is a prevailing "conventional wisdom" that the elderly heal more slowly than the young, there are few control data and animal experiments to support this notion. Nutrition can have a profound effect on the healing of wounds. Prolonged protein starvation can inhibit collagen formation, while high protein diets can enhance the rate of tensile strength gained during wound healing. Local influences that can affect wound healing include infection, inadequate blood supply, and the presence of a foreign body.

Fibrous Tissue Interface. The very presence of the implant provides a dead space in tissue that attracts macrophages to the implant–tissue interface (24). These cells are attracted to the prosthesis as they are to any dead space (e.g., bursa or joint space), presumably because of certain microenvironmental conditions (e.g., low O_2 and high lactate). In this regard it is not clear why macrophages are absent from the surface of osseointegrated implants (see below).

Macrophages along with fibroblasts of the scar comprise synovial tissue (25) that can be considered the chronic inflammatory response to implants (unless the device is apposed by osseous tissue, i.e., osseointegrated). This process is often termed "fibrous encapsulation" (26,27). The presence of regulatory cells such as macrophages at the implant-tissue interface can profoundly influence the host response to a device because these cells can release proinflammatory mediators if irritated by the movement of the device or substances released from the biomaterial (28). The inflammatory response of the synovial tissue around implants is comparable to the inflammation that can occur in the synovium lining any bursa (e.g., bursitis); hence, the response to implants has been termed "implant bursitis" (25).

Response to Implants in Bone: Osseointegration, Bone Ingrowth, Chemical Bonding of Bone to a Biomaterial. Wound healing governs the makeup of the tissue that forms around implants. Because of its capability for regeneration bone should be expected to appose implants in osseous tissue, and form within the pore spaces of porous coatings. Is this bone bonded in any way to the implant? Bonding of a prosthesis to bone would enhance its stability, limiting the relative motion between the implant and bone. In addition, bonding might provide a more favorable distribution of stress to surrounding osseous tissue.

Bonding of bone to an implant can be achieved by mechanical or chemical means., Interdigitation of bone with PMMA bone cement or with irregularities in implant topography, and bone ingrowth into porous surfaces, can yield interfaces capable of supporting shear and tensile as well as compressive forces. These types of mechanical bonding have been extensively investigated and are reasonably well understood. Chemical bonding of bone to materials could result from molecular (e.g., protein) adsorption–bonding to surfaces with subsequent bone cell attachment. This phenomenon has undergone intensive investigation in recent years but is not yet as well understood as mechanical bonding.

The term "osseointegration" has been used to describe the presence of bone on the surface of an implant with no histologically (light microscopy) demonstrable intervening nonosseous (e.g., fibrous) tissue. All implants in bone should become osseointegrated unless the bone regeneration process is inhibited.

The bone ingrowth into a porous-surface coating on an implant leads to an interlocking bond that can serve to stabilize the device. In order for the porous material to accommodate the cellular and extracellular elements of bone, the average pore diameter should be above $\sim$100 μm.

The bone ingrowth process proceeds in two stages. The surgical trauma of implantation initially leads to the regeneration of bone throughout the pores of the coating. Then mechanical stress-induced remodeling leads to resorption of bone from certain regions of the implant and continued formation and remodeling of bone in other regions.

Previous investigations have provided evidence of bone bonding to many different types of calcium phosphate

materials, calcium carbonate substances, and calcium-containing "bioactive" glasses. Chemical bonding was evidenced by the high strength of the implant-bone interface that could not be explained by a mechanical interlocking bond alone. In addition, TEM has shown that there is no identifiable border between these calcium-containing implants and adjacent bone.

Many recent studies have investigated the bonding of bone to one particular calcium phosphate mineral, hydroxyapatite, chosen because its relationship to the primary mineral constituent of bone; natural bone mineral is a calcium-deficient carbonate apatite. Experiments have been performed on both hydroxyapatite coated metallic implants and on particulate and block forms of the mineral employed as bone substitute materials. Histology of specimens from animals and retrieved from human subjects show that a layer of new bone $\sim$100 μm in thickness covers most of the hydroxyapatite surface within a few weeks of implantation and remains indefinitely. This layer of bone is attached to the surrounding osseous tissue by trabecular bridges.

In studying the mechanism of bone bonding, researchers have found that within days of implantation, biological apatites precipitate (from body fluid) onto the surface of the calcium-containing implants. These biological apatites are comparable to the carbonate apatite that is bone mineral. Proteins probably adsorb to this biological mineral layer thereby facilitating bone cell attachment and the production of osteoid directly onto the implant. This osteoid subsequently undergoes mineralization as it does normally in osteogenesis, thus forming a continuum of mineral from the implant to the bone. In this light, the bone cell responds to the biological apatite layer that has formed on the implant and not directly to the implant itself. Recent studies have shown that this biological apatite layer forms on many different calcium phosphate substances, explaining why bone bonding behavior has been reported for many different types of calcium phosphate materials. Of course, the clinical value of this phenomenon will depend (*1*) for coatings, on how well these substances can be bonded to implants; and (*2*) for bone graft substitute materials, their strength, modulus of elasticity, and ability to be resorbed. However, the finding that bone can become chemically bonded to certain biomaterials is a significant advance in our understanding of the implant-bone interface.

Effects of Implant-Induced Alterations of the Mechanical Environment. The presence of the implant can alter the stress distribution in the extracellular matrix (ECM), and thereby reduce or increase the strains experienced by the constituent cells. Many studies have demonstrated immobilization-induced atrophy of certain tissues resulting from the decrease in mechanical strains. Loss of bone mass around stiff femoral stems and femoral condylar prostheses of total hip and knee replacement devices has been associated with the reduced strains due to "stress shielding". Hyperplasia and hypertrophy of tissue in which mechanical strains have increased due to the presence of an implant have also been evidenced.

Criteria for Assessing Acceptability of the Tissue Response

The *in vivo* assessment of tissue compatibility of biomaterials requires that certain criteria be implemented for determining the acceptability of the tissue response relative to the intended application of the material–device. The biomaterial–device should be considered biocompatible only in the context of the criteria used to assess the acceptability of the tissue response. In this regard, every study involving the *in vivo* assessment of tissue compatibility should provide a working definition of biocompatibility. Biomaterials–devices implanted into bone can become apposed by the regenerating osseous tissue, and thus be considered compatible with bone regeneration. However, altered bone remodeling around the device due to stress shielding, with a net loss of bone mass (i.e., osteopenia), could lead to the assessment that the material–device is not compatible with normal bone remodeling. In situations in which the implant is surrounded by fibrous tissue the macrophages on the surface of the material are the expected response to the dead space produced by the very presence of the implant. The synovial tissue thus produced might be considered an acceptable response relative to the chemical compatibility of the material. Utilization of the thickness of the scar capsule around implants alone as a measure of biocompatibility is problematic because it can be influenced by movement of the tissue at the site relative to the implant.

The cellular and molecular make-up of tissue and the interactions among these components are complex. Therefore, criteria for assessing certain features of the biocompatibility of biomaterials–devices should focus on specific aspects of the biological response. The tissue compatibility of materials should be assessed specifically in the context of the effects of the material–device on certain aspects of the response. Moreover, it is important to note that materials yielding acceptable tissue compatibility in one site of implantation might yield unfavorable results in another site.

Degeneration of the Biomaterial–Tissue Interface

As noted earlier it is the wound healing response that initially establishes the tissue characteristics of the implant–tissue interface. Several agents have the potential for initiating degenerative changes in the interface tissue. Others probably act as promoters to stimulate the production of proinflammatory mediators that stimulate tissue degradation, and potentiate the failure process. Of the many factors affecting the implant–tissue interface, motion of the prosthetic component and particulate debris are two of the most important. However, it is difficult to determine the causal relationships between these factors and implant failure from only studying the end-stage tissue. Other histopathological findings and laboratory studies indicate that metal ions and immune reactions might play roles in the degenerative processes leading to prosthesis loosening in certain patients. Systemic diseases and drugs employed for the treatment of the disorders could also serve as factors contributing to the breakdown of the implant–bone interface. Finally, there might be interindividual differences in genetically determined cellular responses that could explain why prostheses fail in

some patients in whom there is a low mechanical risk factor for failure.

Effects of Implant Movement. Movement of the implant relative to the surrounding tissue can interfere with the wound healing response by disrupting the granulation tissue. In the case of implants in bone this relative movement, if excessive, can destroy the stroma required for osseous regeneration, and a fibrous scar results. Another important effect of implant motion is the formation of a bursa within connective tissue in which shearing and tensile movement has led to disruption of tissue continuity and led to the formation of a void space or sack (lined by synovial-like cells). It is to be expected, then, that tissue around prosthetic components removed due to loosening might display features of synovial-like tissue. The presence of synovial cells (macrophage and fibroblast-like cells) is important because they could be activated by other agents, such as particulate debris, to produce proinflammatory molecules. The process of activation of this tissue might be similar to that which occurs in inflammatory joint synovium or bursitis.

An explanation of how prosthetic motion leads to the formation of the synovial-like tissue can be found in previous studies (29) that have shown that "synovial lining is simply an accretion of macrophages and fibroblasts stimulated by mechanical cavitation of connective tissue". These findings are based on experiments in which the mechanical disruption of connective tissue was produced by injection of air and/or fluid into the subcutaneous space of animals (30). The resulting sack was initially described as a "granuloma pouch". Later studies (29) demonstrated that the membrane lining the pouch displayed the characteristics of synovium, and referred to this tissue as "facsimile synovium".

Prosthetic motion can also contribute to wear of the prosthetic component abrading against the bone cement sheath or surrounding bone, thereby generating increased amounts of particulate debris that might contribute to activation of the macrophages and synovial-like cells at the implant–tissue interface.

Effects of Implant-Derived Particles. Particulate debris can be generated from the abrasion of the implant against surrounding tissue. Understandably, the potential for wear is greater with materials–devices rubbing against a hard surface such as bone and with the articulating components of joint replacement prostheses. This particulate debris can induce changes in the tissue around the implants. Adverse responses have been found to both metallic and polymeric particles. The biological reactions to particles are related to (*1*) particle size, (*2*) quantity, (*3*) chemistry, (*4*) topography, and (*5*) shape. While it is not clear what role each of these factors play in the biological response, particle size appears to be particularly important. Particles small enough to be phagocytosed (<10 μm) elicit more of an adverse cellular response than larger particles.

Particulate metallic particles (viz., cobalt–chromium alloy particles) can induce rapid proliferation of macrophages and focal degeneration of synovial tissues (31). Because previous animal investigations and histopathological studies of tissues retrieved from human subjects have suggested that titanium alloy is more "biocompatible" than cobalt–chromium alloys it has been assumed that titanium particulate debris would be less problematic than particles of cobalt–chromium alloy. Histology of pigmented tissue surrounding titanium implants has generally revealed considerably fewer macrophages and multinucleated foreign body giant cells than seen around cobalt–chromium alloy particles and polymeric particulate debris. However, titanium alloy particles generated by the abrasion of femoral stems against bone cement in human subjects can cause histiocytic and lymphoplasmacytic reactions to the metallic particles (32). Titanium particles have also been found to cause fibroblasts in culture to produce elevated levels of PGE_2. These findings show that there may be adverse aspects of the biological response to titanium particles as well as to cobalt–chromium alloy particulate debris.

Many investigations evaluating the histological response to polyethylene and polymethylmethacrylate particles in animals and in tissue recovered from revision surgery have revealed the histiocytic response to these polymer particles. Moreover, it has been shown that this macrophage response can lead to bone resorption.

Synovial cells also respond to calcium-containing ceramic particles (33). Local leukocyte influx, proteinase, PGE_2, and tumor necrosis factor (TNF) levels have been measured after injection of calcium containing ceramic materials into the "air pouch model" described above. The TNF was detected in significant amounts after injection of the ceramics. These substances also provoked elevated leukocyte counts and levels of proteinase and PGE_2, showing that substances with surface chemistries that elicit a beneficial tissue response (e.g., bone bonding) when implanted in bulk form can cause destructive cellular reactions when present in particulate form.

Investigations indicate that most biomaterials, when present in particulate form in a size range small enough to be phagocytosed (<10 μm), can elicit a biological response that could cause the bone resorption that initiates and promotes the loosening process. This degenerative process has been referred to as "small particle disease".

Metallic Ions. Animal and human investigations have revealed elevated levels of metal ions in subjects with certain types of implants (viz., total joint replacement prostheses). Our knowledge is still incomplete with respect to the mechanisms of metal ion release. Results are often variable with respect to the concentration of specific metal ions in certain tissues and fluids. The fact that metal in ionic form is often not distinguished from that present as particles serves to confound interpretation of results.

Rises in serum and urinary chromium levels in patients who have undergo conventional cemented cobalt–chromium alloy hip replacement have previously been reported (34). However, an attempt to determine the valency of chromium as either $+3$ (III) or $+6$ (VI) from the concentration of metal ion in blood clot was not successful. This experiment was based on the fact that erythrocytes display

a unidirectional uptake of Cr(VI) while effectively excluding Cr(III). The distinction of the valency of chromium is important because Cr(VI) is much more biologically active than Cr(III).

Unfortunately, our knowledge of the local and systemic biological and clinical sequelae of metal ion release has not significantly advanced over the past several years. Addition of cobalt ions in the form of cobalt fluoride solutions to the media of synovial cells can stimulate their production of neutral proteinases and collagenase (35). These findings may be relevant to findings of tissue degradation (e.g., osteolysis) around implants in that metal ions could activate synovial cells in the surrounding synovial tissue to produce agents that promote tissue degeneration.

Diseases and Drugs. There has been little work correlating the failure of implants with disease states and drugs employed to treat the disorders. Some observations indicate that antiinflammatory agents, as well as certain anticancer drugs, can reduce the amount of bone formation around devices in the early stages of wound healing after implantation. Little is known, however, about the role of these and other agents on tissue remodeling and degeneration at the biomaterial-tissue interface.

Immune Reactions

It is not infrequent that two patients matched for sex, age, weight, activity level, and other factors, that might be expected to affect the performance of the prosthesis (implanted with the same device by same surgeon), have very different outcomes. This suggests that immune reactions, or genetically determined responses, might play a role in the failure of prostheses in some patients.

Immune responses include antibody and cell-mediated reactions and activation of the complement system. Certain small molecules released from implants (e.g., metal ions), while not antigenic themselves, can bind to existing antibodies and then to larger antigenic molecules or carrier proteins and subsequently elicit antibody production by activation of B lymphocytes by the small molecule (the "hapten") and by activation of helper T lymphocytes by the carrier protein to which it is bound.

The cell types that might be expected to occur at sites of antibody and cell-mediated reactions are not often found in tissue retrieved with revised devices. These cells include lymphocytes and plasma cells. The finding of occasional lymphocytic infiltrates in peri-implant tissue does not provide enough information for the role of immune reactions to implant. Immune reactions to polymeric materials (viz., silicone) have also been suggested as the cause of certain systemic diseases. However, mechanisms for such a response, and its prevalence, remain in question. Much more additional work is necessary to determine the role of immune reactions in the response to implantable devices.

The complement system, comprising circulating proteins and cell-surface receptors, plays an important role in immune processes engaged in the host defense against infectious agents. Te complement system consists of 20–30 proteins circulating in blood plasma. Most of these are inactive until they are cleaved by the chemical action of an enzyme of the interaction with a biomaterial surface. Once activated, the proteins can initiate a cascade of reactions resulting in the mobilization of immune cells resulting in inflammatory processes. Previous studies have demonstrated that many biomaterials can activate (cleave) certain molecules (C_3 and C_5) in the complement system and thereby stimulate the alternative pathway of the immune response. It has been suggested that complement activation by biomaterials could play a role in adverse reactions to certain devices. However, additional studies are required.

One form of cell-mediated immune reaction associated with implants, that has been studied, is the delayed hypersensitivity response. "Metal allergy" has been incriminated as the cause of failure in certain patients (36). However, results obtained to date are not definitive. "The incidence of metal sensitivity in the normal population is high, with up to 15% of the population sensitive to nickel and perhaps up to 25% sensitive to at least one of the common sensitizers Ni, Co, and Cr. The incidence of metal sensitivity reactions requiring premature removal of an orthopedic device is probably small (less than the incidence of infection). Clearly, there are factors not yet understood that caused one patient, but not another, to react" (37).

A similar situation exists with respect to sensitivity reactions to polymeric materials including bone cement (PMMA). The monomer of PMMA is a strong skin sensitizer (38). However, failure of cemented devices has not yet been correlated with a hypersensitivity response in patients.

The fact that there is no clear etiology of the prosthesis loosening in some patients while in other individuals with multiple risk factors for failure the prosthesis functions well has suggested that there may be genetic determinants for loosening.

Carcinogenicity

Chromium and nickel are known carcinogens and cobalt is a suspected carcinogen. Therefore, it is understandable that some concern might be raised about the release of these metal ions into the human body from implants. Fortunately, there have been few reports of neoplasms around implanted devices (e.g., total joint replacement prostheses). While no causal relationship has been evidenced, there is a high enough index of suspicion to warrant serious investigation of this matter through epidemiological and other studies. The use of porous coated metallic devices (with large surface area) in younger patients (e.g., noncemented total joint replacement prostheses) has added to concern about the long-term clinical consequences of metal ion release because of the significant increase in exposure of patients to metal ions.

Prior publications have reviewed the relationship of metallic ion release to oncogenesis (1), and reports of neoplasms found around orthopedic implants have been reviewed (39). The difference in the tumor types, time to appearance, and type of prosthesis confounds attempts to conclude an association of the neoplasm to the implant materials and released moieties.

In an epidemiological investigation conducted in New Zealand (40), >1300 total joint replacement patients were followed to determine the incidence of remote site tumors. The incidences of tumors of the lymphatic and hemapoietic systems were found to be significantly greater than expected in the decade following arthroplasty. It is important to note that the incidences of cancer of the breast, colon, and rectum were significantly less than expected. The investigators acknowledged that while the association might be due, in part, to an effect of the prosthetic implants, other mechanisms, particularly drug therapy, require consideration. Somewhat similar results were obtained from another recent study (41) of the cancer incidence in 443 total hip replacement (McKee–Farrar) patients operated on between 1967 and 1973 (followed to the end of 1981). The risk of leukemias and lymphomas increased while the risk of breast cancer decreased. The authors concluded that the local occurrence of cancer associated with prostheses made of cobalt–chromium–molybdenum as reported in the literature as well as animal experiments indicate that "chrome–cobalt–alloy plays some role in cancerogenesis (sic)".

In a recent publication (42), a nationwide cohort study performed in Sweden to evaluate cancer incidence among > 100,000 hip replacement patients found no overall cancer excess relative to the general population. The standardized incidence ratios (SIRs) were, however, elevated for prostate cancer and melanoma and reduced for stomach cancer risk. Long-term follow-up (> or = 15 years) revealed an excess of multiple myeloma. The study found no material increase in risk for bone or connective tissue cancer. The investigators noted that, while hip implant patients had similar rates of most types of cancer to those in the general population, excesses in certain types of cancers warranted further investigation, particularly because of the ever-increasing use of hip implants at younger ages.

BACTERIAL INFECTION

Biomaterial surfaces can provide favorable substrates for the colonization by bacteria. The adherence of bacteria to solid surfaces is facilitated by the their production of a "biofilm". The biofilm is a complex structure comprising bacterial cells encapsulated in a polymeric matrix. The detailed composition of the matrix has yet to be fully determined. Little is still known about how certain biomaterials may favor the production and adherence of a biofilm. Studies are still seeking to understand how certain material characteristics might favor bacterial colonization.

Infection following material implantation may be defined as multiplication of pathogenic microorganisms in the tissue of the host after a material implantation, causing disease by local cellular injury, secretion of a toxin, or antigen–antibody reaction in the host. The pathogenic microorganisms could be bacteria, fungi or viruses; the mostly common pathogenic microorganisms are bacteria.

Implant-related infection is one of the most serious and difficult complications to treat, often requiring reoperation including the surgical removal of the implant, and it may result in osteomyelitis, amputation, or even death. About 25–50% of infected vascular prostheses for cardiac, abdominal, and extremity vessel replacement cases result in amputation or death (43–45). Infectious complications are the principal concerns in the use and development of implanted materials for several indications.

The Biological Response to Bacterial Infection

We have noted above that the surgical trauma associated with the implantation of medical device is an injury that elicits an inflammatory response. Bacteria are another form of injurious agent that similarly elicits inflammation. While there are similarities in the cellular reactions to these two forms of injury there are important differences. That cells of the immune system are involves in the inflammatory reaction to bacterial infection is cause for use of the term "immune response" to describe the biological process elicited by a bacterial infection.

The immune response (also called the "immune reaction) is a defense function of the body that protects it against invading pathogens, foreign proteins, and malignancies. It consists of the "humoral immune response" and the "cell-mediated immune response". In the humoral immune response, B lymphocytes produce antibodies that react with specific antigens brought by invading pathogens, foreign proteins, and malignancies. The antigen–antibody reactions activate the complement cascade, which causes the lysis of pathogens or cells bearing those antigens. The humoral response may begin immediately on invasion by an antigen in acute type or up to days later in chronic type. In the cell-mediated immune response, T lymphocytes mobilize tissue macrophages in the presence of foreign antigen, which causes the pathogens or cells bearing those antigens been taken by phagocyte.

An implant-related infection occurs when an adequate number of a sufficiently virulent organism overcomes the host's immune response and establishes a focus of infection at the implant site. Implant-related infections remain a formidable challenge to the surgeon as well as material scientist. The high success rate obtained with antibiotic therapy in most bacterial diseases has not been obtained with implant-related infections for several known and as yet unidentified reasons. One important factor is that the sites on and around implants colonized by bacteria have little or no blood supply and thus do not allow blood-borne antimicrobial agents to reach the bacteria. The biofilm in which the bacteria grow may also shield the pathogens from the antimicrobial agents. Illness, malnutrition, and inadequacy of the immune system may be other factors that allow for the development implant-related infections.

Classification of Pathogens in Environment

Pathogens in the environment can be divided into three categories: primary pathogen, opportunistic pathogen, and nonpathogen. Primary pathogens are organisms that can cause infection in normal host when it has attached to the host's tissue and has gained sufficient numbers. It is also called a professional pathogen. Only a very small proportion of microbial species may be considered to be primary or

professional pathogens, and even among these species only a relatively small number of clones have been shown to cause infection. Pathogenic organisms are highly adapted to the pathogenic state and have developed characteristics that enable them to be transmitted, to attach to surfaces, to invade tissue, to escape host defenses, to multiply, and thus to cause infection.

Opportunistic pathogens are those organisms that can only cause infection in impaired hosts. For opportunistic pathogens, the state is the main determinant of whether infection will be the outcome of their interaction with the host's local tissue. This group of organisms may lack effective means to overcome an unimpaired host's defense mechanisms. They have limited growth opportunities outside their restricted niche in an unimpaired individual. As a result, infection may be only a rare consequence of the host–microbe encounter.

Nonpathogens are harmless members of the normal flora in healthy individuals. They may, however, in some rare situations, act as virulent invaders in an individual with severe deficiencies in host defense mechanisms.

The Most Common Bacteria that Cause Implant-Related Infection and Routes of Transmitting Pathogens

Studies of infected implants that have been retrieved for analysis indicate that a few species seem to dominate implant-related infections. Coagulase negative staphylococci are most frequently involved in the implant-related infections. Aerobic Gram-negative bacteria and anaerobic bacteria, which are usually present in deep infections (46), can also cause implant-related infection. *Staphylococcus aureus* (*S. aureus*) and *Staphylococcus epidermidis* (*S. epidermidis*) have been most frequently isolated from infected implant material surfaces. However, *Escherichia coli* (*E. coli*), *Pseudomonas aeruginosa*, β-hemolytic streptococci, and enterococci have also been isolated (43,47). These bacteria more often act as a component of mixed infections.

The different physical and chemical properties of implant material surfaces appear to be responsible for favoring infection with certain bacteria. *Staphylococcus aureus* is mostly involved in infection of metallic implants, such as metallic artificial joints, whereas *S. epidermidis* is a primary cause of infection of polymeric biomaterial implants, such as vascular grafts, catheters, and shunts (47).

The most pathogenic species is *S. aureus* because the infection caused by *S. aureus* often results in a much higher rate of mortality, and it is rarely cleared without removal of the implant. The recent emergence of *S. aureus* that are more resistant to all approved antibiotics raises more serious concerns for the future (48).

Implant-related infections may be the result of bacterial contamination of the implant material prior to its implantation. Pathogenic microorganisms may obtain access into the body by, direct contact, airborne spreading, contaminated water transmission, and blood stream transmission. If microorganisms exit in the host's tissue or on the skin or mucous membrane away from the implant site and break through blood barrier (e.g., associated dental treatment) they can gain access to the blood circulation. This can bring the microorganisms to the implant site (i.e., hematogenous infection).

Risk Factors for Implant-Related Infections

Implant-related infections occur when an adequate number of a sufficiently virulent organism overcomes the host's defense systems. During this process, many factors may be involved or even cooperate in establishing a focus of infection at the local implant area. The risk factors may involve the implant material, the process of implantation surgery, and the host.

Material-Related Factors. A large surface increases the possibility of microorganism attachment and thus can lay a role in implant-related infection. The avascular zone surrounding a device also favors infection as it contains tissue fluid and is often free of microorganism-monitoring agents because of the absence of blood circulation. A small initial number of microorganisms can grow to significant numbers and cause infection without interruption.

Sterile implant materials are commonly packaged in sterile paper, cloth, and plastic bags. However, all of these may be accidentally broken without notice and allow bacterial contamination of the implant.

Implantation-Related Factors. Implantation-related factors include the operating environment, skin and wound care, and surgical technique. In the operating room, airborne microorganisms (usually Gram-positive bacteria) are a source of wound contamination, originating with operating room personnel. Each person in the operating room sheds as many as 5000–55,0000 particles/min. Conventional operating room air may contain 10–15 bacteria/ft^3(49).

The microorganisms present on the host's skin are another source of wound contamination. Although the skin and hair can be sterilized with disinfectant agents, it is almost impossible to sterilize the hair follicles and sebaceous glands because the disinfectants now used in surgery do not penetrate an oily environment. Many disinfectants that do penetrate the oily environment, such as hexachlorophene, are absorbed by the body and have potentially toxic side effects. For this reason, skin preparations now used in surgery have a limited effect on sebaceous glands and hair follicles where microorganisms normally reside and reproduce (49). Because the skin can never be disinfected completely, the number of residual microorganisms present on the skin after disinfection builds the possibility of infection.

Any factor or event that delays wound healing increases the risk of implant-related infection. Ischemic necrosis, seroma, hematoma, wound infection, and suture abscesses are common preceding events for implant-related infection. Surgical technique and operating time also contribute to infection rates.

Host-Related Factors. Systematic factors that can contribute to implant-related infections include: immunological status, nutrition, chronic disease, and infection at a remote site or bacteremia caused by other reasons. A

deficiency in the host's defense mechanisms predisposes the host to infection by specific groups of opportunistic pathogens (49). Deficiencies in the immune system may be acquired (such as acquired immune deficiency syndrome, AIDS) or may result from congenital abnormalities. Malnutrition and chronic disease decreases both the immune and inflammatory response to microorganism invasion. Although the contaminating microorganisms may be few in number, the altered host's defense mechanisms implies that even small bacterial counts have to be regarded as highly virulent species.

If there is infection at a remote site in the host, the microorganisms can be brought to the implant site by blood stream and cause implant-related infection. Under several other conditions which the blood barrier are broken, such as a dental treatment, microorganisms are transported by the blood stream to find their way to the implant site, causing hematogenous infections (50,51). Infection of total hip arthroplasties after dental treatment is not rare (52).

The Most Common Feature of Implant-Related Infection: Biofilm

At the implant site, the surface of the material is immersed in the tissue fluid of the host's local tissue. If microorganisms appear, they have a strong tendency to colonize surfaces to form a microecosystem in which various microbial strains and species grow in a complex community-like structure, which is called biofilm.

Biofilms are defined as bacterial populations reside and produce in matrix that adheres to a surface, interface, or each other. During most implant-related infections, microbial products may assist the development and persistence of the infection in association with adsorbed macromolecules from the biological environment in which the implant material is placed. In the presence of implant material, bacteria elaborate a fibrous exopolysaccharide material called the "glycocalyx". The glycocalyx modifies the local environment in favor of the pathogen by hiding and protecting the organism from surfactants, antibodies, phagocytes, and antimicrobial agents. This increases the population of microorganisms on the surface of implant materials *in vivo* (53,54). These protective biofilms may act as bases in predisposing to tissue invasion and also result in the persistence of infection. Biofilms are implant-associated and troublesome. They have been reported to be 500 times more resistant to antibiotics than planktonic cells (55).

Growth of the organisms is the main mechanism of multiplication in a biofilm and eventually leads to the formation of a thick film. The biofilm is formed in three phases. The first phase is the formation of "conditioning film". As soon as the implantation of material performs, the material surfaces adsorbed macromolecules from the surrounding fluid, forming a conditioning film. The macromolecules are a number of extracellular proteins that interact with host intracellular matrix and blood proteins. For example, joint materials adsorb macromolecules from synovial fluid, bone materials adsorb macromolecules from plasma, while dental materials adsorb macromolecules from salivary fluid. The conditioning film forms within seconds of exposure of the implant to a body fluid (56). This conditioning film provides a suitable substrate for microorganism's adhesion.

The second phase is an initial, reversible adherence of microorganisms to the conditioning film. This adhesion depends on the physicochemical characteristics of the microbial cell surface, the material surface and the conditioning film. Microorganisms can reach the surface via various transport mechanisms, such as diffusion, convection or sedimentation (57). Implants can become contaminated before or during surgery, and more likely, by hematogenous seeding (58). Several factors are reported to contribute to this initial adherence, including surface hydrophobicity, proteinaceous adhesins, and capsular polysaccharides, such as fibrinogen, fibronectin, thrombospondin, von Willebrand factor, collagen, bone sialoprotein, and elastin (59). It seems that different bacteria is helped by different group of factors. *Staphylococcus aureus* appears to be enhanced by mostly fibronectin and plasma glycoprotein in adhering to polymethylmethacrylate *in vivo*, and this may contribute to the establishment of infection (60).

The third phase of biofilm development is microcolony formation and exopolymer production, which results in the firm anchoring of the biofilm and complex biofilm architecture. The adhered organisms multiply and form microcolonies and higher ordered structure glycocalyx. As soon as glycocalyx have been formed, the organisms gain some resistance. In favor of the protection, the microorganisms keep multiplying within this matrix. New layers of film are added and allowing the microorganisms room to multiply, forming thick biofilms (56). Biofilm protects the resident microorganisms against environmental attacks and antibiotics. However, the mechanism for resistance is not well understood.

As biofilms grow thicker and thicker, microorganisms on the periphery of the expanding biofilm may detach, which plays a large part in the pathogenesis of septic processes (61,62).

Latent Infections

As the biofilm protects microorganisms against environmental attacks and antibiotics, the microorganisms can survive in the biofilm for a long period of time when the host's defense system is strong or sufficient concentration of antibiotics is exit. There is no clinical symptom or sign of infection, but the microorganisms exit in the implant site. However, if the host's defense system becomes weakened or a sufficient concentration of antibiotics is no longer administered, the microorganisms may become more active, and cause a latent infection.

Latent infection also can be caused in other situations by a remote wound. The remote wound can give microorganisms access to the blood circulation. The blood stream can bring the microorganisms to the implant, causing a latent infection. Infection of total hip arthroplasty after dental treatment is not rare (52).

The Outcomes of Infection

Implant-related infection produces all the symptoms of infectious inflammation with a wide spectrum of severity.

The clinical presentation is determined largely by the virulence of the infecting pathogen, the extent of the area involved, the location of the infection and the nature of the infected host tissue. The infection may cause large changes to the host's internal environment as well as the implant material. As we mentioned in former section, *S. aureus* is a common pathogen in implant-related infection. In the mean while, it is particularly a very virulent pathogen in this setting and usually produces a fulminant infection.

The early stage of implant-related infection may be obvious or obscure. Signs and symptoms vary with the location and extent of tissue or organ involvement. Common characteristics of infection, such as fever, chills, nausea, vomiting, malaise, erythema, swelling, and tenderness may or may not be present. The classic triad is fever, swelling, and tenderness or pain. Tenderness or pain probably is the most common and earliest symptom. Swelling may be mild. Fever is not always a consistent finding.

During the mid-late stage of infection, the severity of the infection, its specific microorganism, and the particular tissue, site, and material involved all introduce morphologic variations in the basic patterns of acute and chronic infection. The implant-related infection can appear as serous inflammation, fibrinous inflammation, suppurative or purulent infection, abscesses, or more seriously lead septicemia, septic shock, or patient death.

Serous inflammation is marked by the outpouring of a thin fluid derived from the blood serum. Fibrinous inflammation is a fibrinous exudate develops when the vascular leaks are large enough. These two are the clinical appearance of mild infection. Fibrinous exudate may convert to scar tissue if the infection is controlled in this stage (63).

Suppurative or purulent infection is commonly seen in implant-related infection. It is characterized by the production of large amounts of pus or purulent exudate consisting of neutrophils, necrotic cells, and tissue fluid. Certain pyogenic (pus-producing) microorganisms (e.g., staphylococci) produce this localized suppuration.

Abscesses are focal localized collections of purulent inflammatory tissue caused by suppuration buried in a tissue. They are produced by pyogenic bacteria. Abscesses have a central region that appears as a cavity full of pus that consists of necrotic tissue, died white blood cells, bacteria, and material. There is usually a zone of neutrophils around this necrotic focus. Vascular dilation, parenchymal and fibroblastic proliferation occurs outside this region, indicating the beginning of repair. Sometimes, the abscess may become walled off by connective tissue that limits it from further spread (64).

Microorganisms on the periphery of the expanding biofilm may detach or separate. These microorganisms may present in blood and can be confirmed by blood culture, which is called bacteremia. If the bacteria are strong enough to survive in blood and produce toxin, it is called septicemia, which can be life-threatening.

Diagnosis of Implant-Related Infection

The specific diagnosis of implant-related infection is dependent, in large part, upon isolation of the pathogen by aspiration of secreted fluid or by culture of tissue obtained at debridement. However, there are other assessments that can indicate an infection, such as blood count and different morphologic examinations.

Roentgenographic studies are helpful in implant-related infections. Plain roentgenograms show soft tissue swelling, joint space narrowing or widening, bone destruction and non-X-ray transparent implant materials. These roentgenograms can reveal (*1*) abnormal lucencies at the material–host tissue interface, (*2*) bone or periosteal reaction, (*3*) motion of components on stress views, or (*4*) changes in the position of implant materials. If initial roentgenograms are normal in the evaluation of a suspected implant-related infection, other imaging modalities that show soft tissue swelling and loss of normal fat planes about the involved site should be used.

Computed tomography (CT) scanning can help determine the extent of surrounding tissue involvement. Pus within the cavity can cause an increased density on the CT scan. Adjacent soft tissue abscesses also are easily seen. However, the use of CT scan is limited if the implanted material is made of metal.

Magnetic resonance imaging (MRI) can also be implemented for evaluating implant-related infections. The images can reflect the increase in water content resulting from edema in the implant or surrounding tissue due to infection. The MRI detects changes much earlier in the course of disease than roentgenograms, because it shows the condition of the surrounding soft tissue.

Fate of Material During Infection

The biofilm and bacterial modification of the microenvironment around implants may affect the biomaterial in several ways. Biodegradable materials, such as collagen, may degrade faster than expected due the elevated levels of enzymes and the changes in pH. Such implants may collapse before being replaced by host tissue and thus become a component of the local abscess. The low pH often found at sites of infection may also accelerate the corrosion of metallic materials. Implant-related infections indirectly affect implant materials by causing the destruction of surrounding tissue thus contributing to loosening of the implant.

Treatments of Implant-Related Infection

Successful treatment of an implant-related infection often depends on both extensive and meticulous surgical debridement and effective antimicrobial therapy. Debridement should be emphasized because infection often persists despite treatment with systemic antibiotic therapy in the absence of extensive and meticulous debridement of the implant-tissue interface.

It is of paramount importance to confirm the microorganism causing the infection. Distinguishing infection from pure inflammation is also very important because they have some similarities. The timing and selection of bacteria culture are critical. Many implant-related infections are deep seated, and adequate culture specimens are difficult to obtain. In spite of this, every effort should be made to obtain a culture specimen. The preferred specimen in implant-related infections is aspirated fluid. A deep

wound biopsy or a curetted specimen after cleaning the wound is acceptable.

Antibiotic therapy should begin as early as implant-related infection is diagnosed. Treatment with systemic antibiotic therapy can at least prevent the infection from transmission. In the meanwhile, local symptoms and signs of infection should be observed carefully. If there are signs show the infection is not under control, a debridement may be indicated.

Treatment of an implant-related infection at mid–late stage may require both systemic antimicrobial treatment and local surgical treatment. Antibiotic treatment alone sometimes may still be sufficient at this stage, however, it should be performed under careful observation and cannot last long. If there are signs that the infection is not under control, surgery is needed.

Surgery may go hand in hand with antibiotic treatment. The purpose of surgery is clearance of the necrotic tissue with the bacteria and augmentation of the host response. Debridement and irrigation removes necrotic and avascular tissue, bacteria, and harmful bacterial products. It is essential when pus is found on aspiration, signifying an abscess, or when roentgenographic changes indicating pus, necrotic material, and chronic inflammation. If an abscess has formed, removal of the implant is indicated.

Frequently, the only way to treat an infected large implant is to remove it in associate with antibiotic treatment.

If the infection is very severe, septic shock may exist and threaten the patient's life. At this situation, the most important work is antishock and save life. Supported by antishock and antibiotic treatment, removal of implant and open debridement, or even amputation must be performed.

How to Prevent Implant-Related Infection

Recognizing the unique characteristics and outcomes of implant-related infections, the best course is prevention. The close relationship and cooperation of implant designer, manufacturer, and surgeon are necessary for prevention of implant-related infection. The implant must be kept free of bacteria, while the surgeon should evaluate the risk of infection in each patient by considering both host- and surgeon-dependent factors. Simply stated, it is much easier to prevent an implant-related infection than to treat it.

Sterilization Methods. There are several methods for sterilizing implants prior to implantation. The first concern when choosing a sterilization method is the physical and chemical properties of the implant material itself as well as the packaging material required to maintain the implant sterile prior to delivery to the operative site. Autoclaving is the method of choice for the sterilization of metallic or heat-resistant implants. The advantages of autoclaving are efficacy, speed, process simplicity, and no toxic residues. The disadvantages are the relatively high temperature of the process (121 °C) may damage some non-heat resistant implant materials as well as the packaging materials. Thus, most nonmetallic implants and packaging materials cannot be sterilized by this method.

Ethylene oxide (EtO) gas sterilization is a low temperature sterilizing process. It is compatible with a wide range of implant and packaging materials. It is commonly used to sterilize a wide range of medical implants, including surgical sutures, absorbable and nonabsorbable meshes, absorbable bone repair devices, heart valves, and vascular grafts. The advantages of EtO are its efficacy, high penetration ability, and compatibility with a wide range of materials. The main disadvantage is EtO residuals in the sterilized materials.

Exposure to ^{60}Co gamma rays is another widespread sterilizing method. Gamma rays have a high penetrating ability. This method of sterilization is widely used for medical products, such as surgical sutures and drapes, syringes, metallic bone implants, knee and hip prostheses. The advantages of ^{60}Co gamma-ray sterilization are efficacy, speed, process simplicity, no toxic residues. The main disadvantages are the very high costs and incompatibility of some radiation sensitive materials such as the fluoropolymer and polytetrafluoroethylene (PTFE).

Medical implant may also be sterilized with machine-generated accelerated electrons called electron beam sterilization. It has a similar range of applications and material compatibility characteristics as the ^{60}Co process. However, the main disadvantages are short penetration distance. This limits its usage. A unique application for this method is the on-line sterilization of small, thin materials immediately following primary packaging.

Several new technologies are emerging that have potential utility for implant material sterilization, such as gaseous chlorine dioxide, low temperature gas plasma, gaseous ozone, vapor-phase hydrogen peroxide, and machine-generated X rays. Machine-generated X rays have the advantage of a nonisotopic source and penetrating power similar to gamma rays.

Prevention of Operative Contamination, Wound Sepsis Contiguous to the Implant, and Hematogenous Infection. The importance of irrigation during and at the end of surgery has been well documented. The principles of no dead space, no avascular tissue, evacuation of hematomas, and soft tissue coverage should be practiced strictly during implantation surgery. Good surgical technique and minimal operating times also contribute to lowering of infection rates. Prophylactic antibiotics are definitely indicated when implants are involved. New methods are on the way of developing. Direct local delivery of polyclonal human antibodies to abdominal implant sites reduced infection severity and mortality in an animal model of implant-related peritoneal infection (65).

Any bacteremia can induce an implant-related infection by the hematogenous route (51,66,67). Dentogingival infections and manipulations are known causes of streptococcal and anaerobic infections in prostheses (51). Pyogenic skin processes can cause staphylococcal and streptococcal infections of joint replacement. Genitourinary and gastrointestinal tract procedures or infections are associated with Gram-negative bacillary, enterococcal, and anaerobic infections of prostheses (66,68). Twenty-to-forty percent of prosthetic joint infections are caused by the hematogenous route (68).

To prevent hematogenous implant-related infection, any factor that might predispose to infection should be avoid before insertion of the implant material. For elective implant surgery, the patient should be evaluated for the presence of pyogenic dentogingival pathology, skin, and other even very small local infection. Perioperative antibiotic prophylaxis is also very important. It has been reported to reduce infections in total joint replacement surgery (69).

For patients with indwelling implant materials, it is of paramount important to diagnose and treat infections in any location as early as possible. This reduce the risk of seeding bacteria to the implant materials hematogenously in large extent. Any situation likely to cause bacteremia, even a dental care should be avoided or seriously evaluated.

Antimicrobial Biomaterials

Two factors necessary for an implant-related infection are attachment of bacteria to the implant surface and multiplication of bacteria to a significant number. Antibacteria materials are designed to decrease the surface attachment of microorganism and/or inhibit the multiplying of microorganism.

A variety of antibiotics incorporating implant materials were designed to inhibit the microorganism against multiplying. Antibiotics for incorporation in materials should have a broad antibacterial spectrum, sufficient bactericidal activity, high specific antibacterial potency, low rate of primary resistant pathogens, minimal development of resistance during therapy, low protein binding, low sensitizing potential, marked water solubility, and stable (70). During these years, various antibiotics have been evaluated both *in vitro* and *in vivo*. The studies are most regarding their suitability for incorporation in materials.

Cefazolin loaded bone matrix gelatin (C-BMG) was made from putting cefazolin into BMG by vacuum adsorption and freeze-drying techniques (71). It was tested for repair of long segmental bone defects and preventing infection in animal experiment. The effective inhibition time to staphylococcus aureus of C-BMG was 22 days in vitro, while 14 days in vivo. The drug concentration in local tissues (bone and muscle) were higher than that of plasma, and the drug concentration in local tissues was higher in early stage, later it kept stable low drug release.

Rifampin-bonded gelatin-sealed polyester was tested in another animal experiment (72). Their results indicated that rifampin-bonded gelatin-sealed polyester grafts were significantly more resistant to bacteremic infection than were silver/collagen-coated polyester grafts.

For developing antibiotics delivery biodegradable materials, polylactide–polyglycolide copolymers were mixed with vancomycin (73). The mixture was compressed and sintered at 55 °C to form beads of different sizes. The biodegradable material released high concentrations of antibiotic *in vitro* for the period of time needed to treat infection. The diameter of the sample inhibition zone ranged from 6.5 to 10 mm, which is equivalent to 12.5–100% of relative activity. By changing the processing parameters, the release rate of the beads was able to be controlled. This provides advantages of meeting the specific requirement for prevention of implant-related infection.

Besides incorporating antibiotics in implant materials, antimicrobial materials have been made in different ways. Bovine serum albumin was used to coat material surfaces by using carbodiimide, a cross-linking agent (74). The inhibition rate of the albumin coating on bacterial adherence remained high throughout the experiment. This suggests the potential use of this cross-linked albumin coating to reduce bacterial adherence and thus the subsequent possibility of prosthetic or implant infection *in vivo*.

In the future, the use of implant materials will surely increase with growing demands for a higher quality of life. In 1997, operating expenses allocated to tissue engineering exceed $450 million and fund the activities of nearly 2500 scientists and support personnel. Growth rate is 22.5% / annum (75). At the beginning of 2001, operating expenses allocated to tissue engineering exceed $600 million and fund the activities of nearly 3300 scientists and support personnel. Spending by tissue engineering firms has been growing at a compound annual rateof 16% (76). However, implant-related infection remains a significant problem in this field. Research on the development of biomaterial surfaces with antimicrobial properties has increased to an annual expenditure of ~$430 million (75).

BIBLIOGRAPHY

1. Black J. Metallic ion release and its relationship to oncogenesis. In: Fitzgerald RHJ, editor. The Hip, Proceedings of the Thirteenth Open Scientific Meeting of the Hip Society. St. Louis: C.V. Mosby; 1985. p 119–213.
2. Bartolozzi A, Black J. Chromium concentrations in serum, blood clot and urine from patients following total hip arthroplasty. Biomaterials 1985;6:2–8.
3. Takamura K, Hayashi K, Ishinishi N, Yamada T, Sugioka Y. Evaluation of carcinogenicity and chronic toxicity associated with orthopedic implants in mice. J Biomed Mater Res 1994;28:583–589.
4. Bouchard PR et al. Carcinogenicity of CoCrMo (F-75) implants in the rat. J Biomed Mater Res 1996;32:37–44.
5. Lombardi AV, Mallory TH, Vaughn BK, Drouillard P. Aseptic loosening in total hip arthroplasty secondary to osteolysis induced by wear debris from titanium-alloy modular femoral heads. J Bone Jt Surg 1989;71A:1337.
6. Blaine TA et al. Increased levels of tumor necrosis factor-a and interleukin-6 protein and messenger RNA in human peripheral blood monocytes due to titanium particles. J Bone Jt Surg 1996;78-A:1181–1192.
7. Gonzales JB, Purdon MA, Horowitz SM. In vitro studies on the role of titanium in aseptic loosening. Clin Orthop 1996;330:244–250.
8. Goodman SB, Fornasier VL, Lee J, Kei J. The effects of bulk versus particulate titanium and cobalt chrome alloy implanted into the rabbit tibia. JBMR 1990;24:1539–1549.
9. Donkerwolcke M, Burny F, Muster D. Tissues and bone adhesives—historical aspects. Biomaterials 1998;19:1461–1466.
10. Nivbrant B, Karrholm J, Rohrl S, Hassander H, Wesslen B. Bone cement with reduced proportion of monomer in total hip arthroplasty: preclinical evaluation and randomized study of 47 cases with 5 years' follow-up. Acta Orthop Scand 2001;72:572–584.

11. de la Torre B et al. Biocompatibility and other properties of acrylic bone cements prepared with antiseptic activators. J Biomed Mater Res 2003;66B:502–513.
12. Thomson LA, Law FC, James KH, Matthew CA, Rushton N. Biocompatibility of particulate polymethylmethacrylate bone cements: a comparative study *in vitro* and *in vivo*. Biomaterials 1992;13:811–818.
13. Davis RG, Goodman SB, Smith RL, Lerman JA, Williams RJ., III The effects of bone cement powder on human adherent monocytes/macrophages *in vitro*. J Biomed Mater Res 1993;27: 1039–1046.
14. Jones LC, Hungerford DS. Cement Disease. Clin Orthop Rel Res 1987;225:192–206.
15. LeVier RR, Harrison MC, Cook RR, Lane TH. What is silicone? Plas Reconstr Surg 1992;92:163–167.
16. Bobyn JD, Spector M. Polyethylene. In: Encycolpedia of Materials Science and Engineering. New York: Pergamon Press; 1987. p 3649.
17. Li S, Burstein AH. Ultra-high molecular weight polyethylene. J Bone Jt Surg 1994;76-A:1080–1090.
18. Schmalzried TP, Jasty M, Harris WH. Periprosthetic bone loss in total hip arthroplasty: polyethylene wear debris and the concept of the effective joint space. J Bone Jt Surg 1992;74-A:849–863.
19. Green TR, Fisher J, Stone M, Wroblewski BM, Ingham E. Polyethylene particles of a 'critical size' are necessary for the induction of cytokines by macrophages in vitro. Biomaterials 1998;19:2297–2302.
20. Vert M, Pascal C, Chabot F, Leray J. Bioresorbable plastic materials for bone surgery. In: Hastings GW, Ducheyne P, editors. Macromolecular Biomaterials. Vol. Chap. 5 Boca Raton: CRC Press, Inc.; 1984. p 119–142.
21. Yannas IV. Natural materials. In: Ratner BD, Hoffman AS, Schoen FJ, Lemons JE, editors. Biomaterials Science: An Introduction to Materials in medicine. San Diego: Academic Press; 1992. p 84–94.
22. Spector M, Lalor PA. *In vivo* assessment of tissue compatibility. In: Ratner BD, Hoffman AS, Schoen FJ, Lemons JE, editors. Biomaterials Science: An Introductory Text. San Diego; CA: Academic Press; 1996. p 220–227.
23. Spector M, Cease C, Xia T-L. The local tissue response to biomaterials. CRC Crit Rev Biocompat 1989;5:269–295.
24. Silver IA. The physiology of wound healing. In: Hunt TK, editor. Wound Healing and Wound Infection. New York: Appleton-Century-Crofts; 1984. p 11–28.
25. Spector M. et al. Synovium-like tissue from loose joint replacement prostheses: comparision of human material with a canine model. Sem Arthr Rheum 1992;21:335–344.
26. Coleman DL, King RN, Andrade JD. The foreign body reaction: a chronic inflammatory response. J Biomed Mater Res 1974;8:199–211.
27. Laing PG, Ferguson AB, Hodge ES. Tissue reaction in rabbit muscle exposed to metallic implants. J Biomed Mater Res 1967;1:135–149.
28. Anderson JM, Miller K. M. Biomaterial biocompatibility and the macrophage. Biomaterials 1984;5:5–10.
29. Edwards JCW, Sedgwick AD, Willoughby DA. The formation of a structure with the features of synovial lining by subcutaneous injection of air: an in vivo tissue culture system. J Pathol 1981;134:147–156.
30. Selye H. Use of "granuloma pouch" technic in the study of antiphlogistic corticoids*. Proc Soc Exp Biol Med 1953;82: 328–333.
31. Howie DW, V-Roberts B. The synovial response to intraarticular cobalt-chrome wear particles. Clin Orthop 1988;232: 244–254.
32. Agins HJ et al., Metallic wear in failed titanium-alloy total hip replacements. J Bone Jt Surg 1988;70A:347–356.
33. Nagase M, Baker DG, Schumacher, Jr., HR. Prolonged inflammatory reactions induced by artificial ceramics in the rat air pouch model. J Rheum 1988;15:1334–1338.
34. Bartolozzi A, Black J. Chromium concentrations in serum, blood clot and urine from patients following total hip arthoplasty. Biomaterials 1985;6:2–8.
35. Ferguson GM, Watanabe S, Georgescu HI, Evans CH. The synovial production of collagenase and chondrocyte activating factors in response to cobalt. J Orth Res 1988;6:525–530.
36. Merritt K, Rodrigo JJ. Immune response to synthetic materials. Clin Orthop 1996;326:71–79.
37. Merritt K, Brown SA. Biological effects of corrosion products from metals. In: Fraker A, editor. Vol. STP 859 Corrosion and Degradation of Implant Material. Philadelphia: American Society for Testing and Materials; 1985. p 195–207.
38. Merritt K. Role of medical materials, both in implant and surface applications, in immune response and in resistance to infection. Biomaterials 1984;5:47–53.
39. Martin A, Bauer TW, Manley MT, Marks KE. Osteosarcoma at the site of total hip replacement. JB J S 1988;70A:1561–1567.
40. Gillespie WJ, Frampton CMA, Henderson RJ, Ryan PM. The incidence of cancer following total hip replacement. JB J S 1988;70B:539–542.
41. Visuri T, Koskenvuo M. Cancer risk after McKee-Farrar total hip replacement. Acta Orthop Scand 1989;60:25.
42. Signorello LB et al. Nationwide study of cancer risk among hip replacement patients in Sweden. J Natl Cancer Inst 2001;93:1405–1410.
43. Gristina AG. Implant-associated infection. In: Ratner BD, Schoen FJ, Lemons JE, editors. Biomaterials science: an introduction to materials in medicine. San Diego: Academic Press; 1996.
44. Gristina AG, Oga M, Webb LX, Hobgood CD. Adherent bacterial colonization in the pathogenesis of osteomyelitis. Science 1985;228:990–993.
45. Gristina AG, Costerton JW. Bacterial adherence to biomaterials and tissue. The significance of its role in clinical sepsis. J Bone Joint Surg Am 1985;67:264–273.
46. Fitzgerald RH, Jr., Infected Total Hip Arthroplasty: Diagnosis and Treatment. J Am Acad Orthop Surg 1995;3:249–262.
47. Gristina AG. Biomaterial-centered infection: microbial adhesion versus tissue integration. Science 1987;237: 1588–1595.
48. Proctor RA. Toward an understanding of biomaterial infections: a complex interplay between the host and bacteria. J Lab Clin Med 2000;135:14–15.
49. William C. General Principles of Infection. Campbell's Operative Orthopaedics. Mosby, Inc.; 1998.
50. Stinchfield FE, Bigliani LU, Neu HC, Goss TP, Foster CR. Late hematogenous infection of total joint replacement. J Bone Joint Surg Am 1980;62:1345–1350.
51. Lindqvist C, Slatis P. Dental bacteremia–a neglected cause of arthroplasty infections? Three hip cases. Acta Orthop Scand 1985;56:506–508.
52. LaPorte DM, Waldman BJ, Mont MA, Hungerford DS. Infections associated with dental procedures in total hip arthroplasty. J Bone Joint Surg Br 1999;81:56–59.
53. Costerton JW, Irvin RT, Cheng KJ. The bacterial glycocalyx in nature and disease. Annu Rev Microbiol 1981;35:299–324.
54. Gristina AG, Kolkin J. Current concepts review. Total joint replacement and sepsis. J Bone Joint Surg Am 1983;65:128–134.
55. Costerton JW, Lewandowski Z, Caldwell DE, Korber DR, Lappin-Scott HM. Microbial biofilms. Annu Rev Microbiol 1995;49:711–745.

56. Barry D. Infected orthopedic prostheses. Infections associated with indwelling medical devices. Washington, DC: ASM Press; 1994.
57. van Loosdrecht MC, Lyklema J, Norde W, Zehnder AJ. Influence of interfaces on microbial activity. Microbiol Rev 1990;54:75–87.
58. Schmalzried TP, Amstutz HC, Au MK, Dorey FJ. Etiology of deep sepsis in total hip arthroplasty. The significance of hematogenous and recurrent infections. Clin Orthop 1992; 200–207.
59. Patti JM, Allen BL, McGavin MJ, Hook M. MSCRAMM-mediated adherence of microorganisms to host tissues. Annu Rev Microbiol 1994;48:585–617.
60. Vaudaux P, Suzuki R, Waldvogel FA, Morgenthaler JJ, Nydegger UE. Foreign body infection: role of fibronectin as a ligand for the adherence of Staphylococcus aureus. J Infect Dis 1984;150:546–553.
61. Neu TR, Marshall KC. Bacterial polymers: physicochemical aspects of their interactions at interfaces. J Biomater Appl 1990;5:107–133.
62. Busscher HJ, Van der Mei HC. Relative importance of surface-free energy as a measure of hydrophobicity in bacterial adhesion to solid surfaces. In: Doyle RJ, Rosenberg M, editors. Microbial cell surface hydrophobicity. Washington DC: American Society for Microbiology; 1990. p 335–359.
63. Mandell B, Dolin. Principles and Practice of Infectious Disease. Churchill Livingstone; 2000.
64. Kuma C. Robbins Patholog. Basis Disease, 1999.
65. Poelstra KA et al. Prophylactic treatment of gram-positive and gram-negative abdominal implant infections using locally delivered polyclonal antibodies. J Biomed Mater Res 2002;60:206–215.
66. Ahlberg A, Carlsson AS, Lindberg L. Hematogenous infection in total joint replacement. Clin Orthop 1978; 69–75.
67. Lattimer GL, Keblish PA, Dickson TB, Jr., Vernick CG, Finnegan WJ. Hematogenous infection in total joint replacement. Recommendations for prophylactic antibiotics. JAMA 1979;242:2213–2214.
68. Inman RD, Gallegos KV, Brause BD, Redecha PB, Christian CL. Clinical and microbial features of prosthetic joint infection. Am J Med 1984;77:47–53.
69. Norden CW. A critical review of antibiotic prophylaxis in orthopedic surgery. Rev Infect Dis 1983;5:928–932.
70. Wahlig H, Dingeldein E. Antibiotics and bone cements. Experimental and clinical long-term observations. Acta Orthop Scand 1980;51:49–56.
71. You HB, Chen AM. The effect of cefazolin loaded bone matrix gelatin on repairing large segmental bone defects and preventing infection after operation, Zhongguo Xiu Fu Chong Jian Wai Ke Za Zhi 2000;14:162–165.
72. Goeau-Brissonniere O.A. et al. Comparison of the resistance to infection of rifampinbonded gelatin-sealed and silver/collagen-coated polyester prostheses. J Vasc Surg 2002;35: 1260–1263.
73. Liu SJ et al. In vitro elution of vancomycin from biodegradable beads. J Biomed Mater Res 1999;48:613–620.
74. An YH et al. Prevention of bacterial adherence to implant surfaces with a crosslinked albumin coating in vitro. J Orthop Res 1996;14:846–849.
75. Lysaght MJ, Nguy NA, Sullivan K. An economic survey of the emerging tissue engineering industry. Tissue Eng 1998;4: 231–238.
76. Lysaght MJ, Reyes J. The growth of tissue engineering. Tissue Eng 2001;7:485–493.

See also ALLOYS, SHAPE MEMORY; POLYMERIC MATERIALS; POROUS MATERIALS FOR BIOLOGICAL APPLICATIONS; RESIN-BASED COMPOSITES.

BIOELECTRODES

ERIC MCADAMS
University of Ulster at Jordanstown
Newtownabbey,
Ireland

INTRODUCTION

Biomedical electrodes are used in various forms in a wide range of biomedical applications, including:

1. The detection of bioelectric events such as the electrocardiogram (ECG).
2. The application of therapeutic impulses to the body [e.g., cardiac pacing and defibrillation and transcutaneous electrical nerve stimulation (TENS)].
3. The application of electrical potentials in order to facilitate the transdermal delivery of ionized molecules for local and systemic therapeutic effect (iontophoresis).
4. The alternating current (ac) impedance characterization of body tissues.

Good electrode design is not as simple and straightforward a matter as is often assumed, and all electrode designs are not equal in performance (1). One must, therefore, not simply choose an electrode with as conductive a metal plate as possible, which unfortunately, was and appears to still be the case in many designs. Probably due to this mistaken view, it would appear that the associated electronic systems are often first developed and the electrode design is left to the end, almost as an afterthought. If the clinician is to properly diagnose the patient's cardiac problem, for example, it is imperative that the measured biosignal is clear, undistorted, and artefict-free. Unfortunately, monitoring bioelectrodes, if they are not chosen correctly, give rise to significant problems that make biosignal analysis difficult, if not impossible. Similarly, stimulation electrodes must be well-chosen if they are to optimally supply the therapeutic waveforms without causing trauma to the patient.

Current or charge is carried by ions inside the patient's body and by electrons in the electronic device itself and in its leads. The "charge-transfer" mechanism between current/charge carriers takes place at the electrode-patient interface and is of major importance in the design of an optimal electrode. Both the electrode–electrolyte interface and the skin under the electrode (collectively known as the contact) give rise to potentials and impedances that can distort the measured biosignal or adversely affect the electrotherapeutic procedure.

Implanted electrodes are generally made from inert or noble materials that do not react with surrounding tissues. Unfortunately, as a consequence, they tend to give rise to large interface impedances and unstable potentials. Implanted biosignal monitoring electrodes, in particular, require stable potentials and low interface impedances to minimize biosignal distortion and artifact problems. External biosignal-monitoring electrodes can generally use

high electrical performance nonnoble materials such as silver–silver chloride without fear of biocompatibility problems (2). They do, however, have to address the additional and very significant problem of the skin with its sizeable impedance and unstable potential. Along with the desired biosignal, one amplifies the difference between the two contact potentials. If the contact potentials were identical (highly improbable), they would cancel each other out due to the use of a differential amplifier. If the potential mismatch were very large (several hundred mV), the amplifier would not be able to cope and would saturate. If the mismatch in contact potentials is small and stable, this mismatch will be amplified along with the biosignal, and the biosignal will appear shifted up or down on the oscilloscope screen or printout paper, which would generally not be a major problem as the additional voltage offset can be easily removed. What is a significant problem, however, is when the contact potentials fluctuate with time. Their mismatch, therefore, varies and the baseline of the biosignal is no longer constant, which leads to the problem termed baseline wonder or baseline drift, which makes analysis of some of the key features of the biosignal difficult. Filtering out of the drift is often not an option, as the filtering often also removes key components of the desired biosignal.

Large mismatched contact impedances can cause signal attenuation, filtering, distortion, and interference in biosignal monitoring. If contact impedances are significant compared with the input impedance of the amplifier, they can give rise to signal attenuation as a result of the voltage divider effect. Attenuation of the signal is not a major problem, after all, the amplifier is going to be used to amplify the signal by a factor of around 1000 (in the case of an ECG). A significant problem develops, however, because the contact impedance varies with frequency. The frequency-dependence of the contact impedance is a consequence of the presence of parallel capacitances at the electrode-electrolyte interface or at the skin under the electrode. At very high frequencies, the contact impedances are very small and, therefore, no attenuation of the high frequency parts of the biosignal exists. At low frequencies, the contact impedances can be very large and, hence, significant attenuation of low frequency components of the biosignal can exist. The overall signal is not only attenuated, it is also distorted with its low frequency components selectively reduced. The measurement system in effect acts as a high pass filter and the signal is differentiated. In the case of the ECG, the P, S, and T waves are deformed, leading in particular to a modification of the S–T segment. The S–T segment is of vital importance to the electrocardiologist, hence the importance of avoiding such biosignal distortions.

50/60 Hz interference can be amplified along with any monitored biosignal due to the mismatch of the contact impedances. Displacement currents flow from power lines through the air to the monitor cables and then through the electrodes and the patient to ground. If the contact impedances are not identical, the displacement currents flowing through the two contact impedances connected to a differential amplifier will give rise to different voltages at the amplifier's inputs. This 50 Hz offset voltage will be amplified along with the desired biosignal and its amplitude is proportional to an electrode–skin impedance mismatch (3).

Other applications, such as electrical impedance plethysmography and electrical impedance tomography (4), do not monitor intrinsic biosignals emanating from the body, but inject small currents or voltages into the body and record the resultant voltages or currents. The electrical properties of the body or a body segment can then be calculated. In many of these applications, the magnitude and mismatch of contact impedances can give rise to significant errors or artifacts (5). As relatively high frequencies are often involved in these techniques, even the series resistance of the gel pad (which is generally ignored) may become significant.

Although interface impedance and potential are generally less critical for implanted stimulation electrodes, many such electrodes (e.g., implanted pacing electrodes) are used to monitor biosignals as well as to deliver the required stimulation impulses. Even in the case of a purely stimulating electrode, a low interface impedance is required to minimize energy waste and to prolong the life of the power source. Various techniques are therefore used to effectively decrease the otherwise large interface impedances of the noble or inert materials used for their biocompatible properties. Electrode material and high electrical performance is generally less critical for external stimulation electrodes such as TENS and external defibrillation electrodes. Current density distribution is of major importance in these applications in order to avoid electrical hotspots and resultant burns to the skin. In some applications, such as TENS and external pacing, it is even sometimes advantageous to use a relatively resistive electrode material or gel, as this has been found to optimize current density distribution under the stimulation electrode.

As in the above applications, the avoidance of current density hotspots is one of several key factors in iontophoretic, transdermal delivery (6). An additional important constraint that is generally not relevant in other electrotherapies is the maintenance of the delicate electrochemical balance at the electrode/reservoir/skin interface. The electrode potential and impedance, as well as the composition of the drug reservoir, must generally remain within certain narrow ranges in order to avoid the deterioration of the electrode, the contamination of the drug reservoir, and the irritation of the patient's skin.

The electrical properties of the electrode contacts are, therefore, of great importance in most applications. Ideally, the contact with the patient should give rise to the following:

- Zero potential. Unfortunately, zero potential is not possible and a more realistic goal is to achieve a low, stable potential at each of the contacts.
- Zero Impedance. Unfortunately, zero impedance too is not possible and a more realistic goal is to achieve impedances at the two contacts that are low and as similar as possible.

The potentials and impedances of the electrode–electrolyte interface and the skin will therefore be studied in more depth in the following sections.

ELECTRICAL PROPERTIES OF ELECTRODE–SKIN INTERFACE

As briefly outlined above, the electrode–electrolyte interface and the skin under the electrode both give rise to potentials and impedances that can either distort any measured biosignal or give rise to problems during electrical stimulation.

The Electrode–Electrolyte Interface

The Electrode–Electrolyte Potential. When a metallic electrode comes in contact with an electrolyte (in body tissues or in an electrode gel), an ion–electron exchange occurs as a result of an electrochemical reaction. A tendency exists for metal atoms M to lose n electrons and pass into the electrolyte as metal ions, M^{+n}, causing the electrode to become negatively charged with respect to the electrolyte (Fig. 1). Reaction (**1**) is termed oxidation.

$$M \rightarrow M^{+n} + ne^- \qquad (\mathbf{1})$$

Similarly, under equilibrium conditions, some of the ions in solution M^{+n} take n electrons from the metal and deposit onto the electrode as metal atoms M. The electrode becomes positively charged with respect to the electrolyte. Reaction (**2**) is termed reduction.

$$M^{+n} + ne^- \rightarrow M \qquad (\mathbf{2})$$

The overall chemical reaction taking place at the interface is therefore

$$M \rightleftharpoons M^{n+} + ne^- \qquad (\mathbf{3})$$

Under equilibrium conditions, the rate at which metal atoms lose electrons and pass into solution is exactly balanced by the rate at which metal ions in solution deposit onto the electrode as metal atoms. The current flowing in one direction, i_0, is equal to and cancels out the current flowing in the opposite direction. The electrode is said to be behaving reversibly and the common value of currents, io, is termed the exchange current (density). Although the net current flowing through the electrode interface is zero, a potential difference is found to exist between the electrode and the electrolyte and depends on the position of the equilibrium between the two processes (**1**) and (**2**). Generally, the metal is negative relative to the electrolyte. The potential difference depends on the relative activities (or concentrations) of the ions present and on the electrode metal (7). This potential has been termed the equilibrium, reversible, or half-cell (i.e., one interface only) potential in the literature.

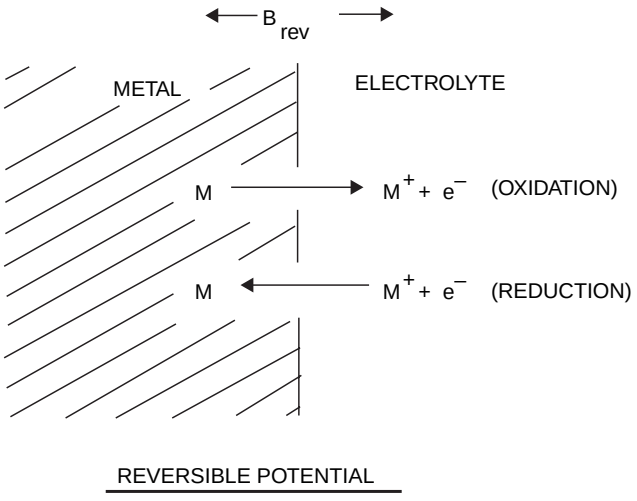


Figure 1. The electrode–electrolyte interface and reactions involved in generating its reversible or equilibrium potential.

When trying to measure the potential of a half-cell (i.e., one interface only), one is immediately faced with a problem, as one requires two electrodes to make a potential measurement, thus effectively creating an electrochemical cell with two electrode–electrolyte interfaces. One, therefore, measures not only the potential of the electrode–electrolyte interface under study, but also that of the second electrode used to complete the circuit. If one uses the same metal for the second electrode to that used in the first, the potentials will be identical (in theory at least) and will cancel each other out. The measured potential will be (theoretically) equal to zero. (In practice, however, slight differences in the composition of the metal used, the electrode surfaces, and in the gel will result in differences in the two half-cell potentials.) If, on the other hand, one uses a different metal for the second electrode, the measured potential of the cell will be due to the combination of the potentials of the two half-cells. It will be impossible to separate the potential of the half-cell under investigation.

In order to resolve this problem, early electrochemists decided to measure all electrode interface potentials with respect to a standardized electrode or reference electrode. The standard hydrogen electrode (SHE) was chosen to be the universal reference electrode and its half-cell potential was specified as zero. Other metal-to-ion interface potentials were then measured with reference to SHE and the entire measured offset voltage was attributed to electrode system being tested.

Hydrogen electrode consists of a platinized plate submerged to one-half its height HCl over which hydrogen gas at atm is bubbled. The half-cell potential of SHE depends on concentration of hydrogen ions in the solution, hence it is quite stable and reproducible. At the time that this decision was reached, the necessary glass blowing and silver soldering were common skills and the SHE was thus easy and inexpensive to make. Although, however, it is no longer convenient for modern routine measurements as a reference electrode (the flowing hydrogen gas is potentially explosive), electrode potentials are standardized with respect to the SHE (7).

The reversible, equilibrium, or half-cell potential of a given electrode–electrolyte interface depends on the activity (almost synonymous with concentration) of the ions taking part in the reactions (Table 1). This potential, E_{rev}, is given by the Nernst equation,

$$E_{rev} = E_0 + [RT/nF]\text{In}[\text{activity of oxidized form/activity of reduced form}] \qquad (1)$$

Table 1. Reversible Potentials for Common Electrode Materials at 25 °C[a]

Metal and Reaction	Potential E^V,V
$Al \rightarrow Al^{3+} + 3e^-$	−1.706
$Zn \rightarrow Zn^{2+} + 2e^-$	−0.763
$Cr \rightarrow Cr^{3+} + 3e^-$	−0.744
$Fe \rightarrow Fe^{2+} + 2e^-$	−0.409
$Cd \rightarrow Cd^{2+} + 2e^-$	−0.401
$Ni \rightarrow Ni^{2+} + 2e^-$	−0.230
$Pb \rightarrow Pb^{2+} + 2e^-$	−0.126
$H_2 \rightarrow 2H^+ + 2e^-$	0.000 by definition
$Ag + Cl^- \rightarrow AgCl + e^-$	+0.223
$2\,Hg + 2Cl^- \rightarrow Hg_2Cl_2 + 2e^-$	+0.268
$Cu \rightarrow Cu^{2+} + 2e^-$	+0.340
$Cu \rightarrow Cu^+ + e^-$	+0.522
$Ag \rightarrow Ag^+ + e^-$	+0.799
$Au \rightarrow Au^{2+} + 3e^-$	+1.420
$Au \rightarrow Au^+ + e^-$	+1.680

[a]The metal undergoing the reaction shown has the magnitude and polarity of standard half-cell potential, E_0. Listed when the metal is referenced to the standard hydrogen electrode (3).

E_{rev} is the reversible, equilibrium, or half-cell potential
E_0 is the standard half-cell potential (measured relative to the standard hydrogen electrode)
R the universal gas constant,
n the number of electrons involved in reaction,
T the absolute temperature (K).

Activity, $a = \gamma C$, where C is concentration and γ, the activity coefficient, is a measure of the interaction between ions. When solution is infinitely dilute, $\gamma = 1$ and activity is equal to concentration.

Note the two components of E_{rev}. One is constant, E_0, wheras the other will vary due to slight variations in concentration, from one electrode to another. If two chemically identical electrodes make contact with the same electrolyte/body, the two interfaces should, in theory, develop identical half-cell potentials. When connected to a differential amplifier, the half-cell potentials of such electrodes would cancel each other out and the offset voltage would be zero. The electrode potentials would, therefore, make zero contribution to a biosignal they were being used to detect. Unfortunately, slight differences in electrode metal or gel result in the creation of offset voltages, which can greatly exceed the physiological variable to be measured. Generally, a more significant problem is that the electrode offset voltage can fluctuate with time, thus distorting the monitored biosignal (8).

The Electrode–Electrolyte Impedance. It has already been stated that in the electrode and the connecting lead, electrical charge is carried by electrons, whereas in the gel and in the human body, charge is carried by ions. A transition exists at the interface between the electrode and the electrolyte where charge is transferred from one kind of carrier to the other. In order for some of the ions in the electrode gel or in the body fluids to transfer their charge across the interface, many must first diffuse to

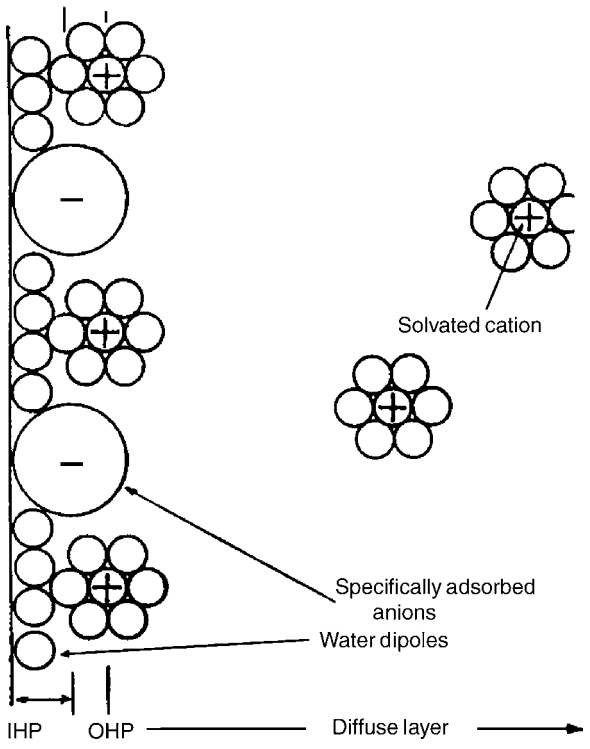


Figure 2. The electrode–electrolyte interface. A metal in an electrolyte forms a double layer of charge. (Redrawn from Ref. 7).

the electrode–electrolyte interface under the influence of electrostatic attraction. Here they stick (or adsorb, as it is termed in electrochemistry) to the electrode surface and form the outer Helmholtz plane (OHP) (7). If the electrode has a negative charge relative to the electrolyte, positive ions will be attracted to the interface region and adsorb onto the electrode surface. As a consequence, there is a layer of negative charge on the metal surface and a layer of equal but opposite charge on the electrolyte side of the interface, both separated by a small distance across the OHP (see Fig. 2). A double layer of charge therefore exists at the interface, and such a system behaves like a parallel-plate capacitor. Not altogether surprising, the interface's capacitance is often termed the double-layer capacitance, C_{dl}, and is connected in parallel to the charge-transfer resistance in our simple equivalent circuit model.

Just in case one believed that the electrode interface was that simple, one must point out, for example, that as well as the cations electrostatically attracted to the negatively charged electrode surface (coulombic adsorption) anions may exits that are adsorbed on the electrode surface and form the inner Helmholtz layer or plane (IHP). These anions have tended to lose their hydration sphere and, consequently, are in close contact with the electrode. As they are negative ions adsorbed onto a negative electrode surface, electrostatic forces cannot be responsible. Some force *specific* to the ion (rather than its electric charge) must be responsible, hence the use of the term specific

adsorption to describe this phenomenon. The van der Waals or chemical forces is thought to be responsible (7).

In order to understand some aspects of the double-layer capacitance, it is good to consider the basic equation for a parallel-plate capacitor. If two identical conductive plates, each of area A cm^2, are separated by a distance d cm, which is filled with a material of dielectric constant ε_0, then the capacitance of this parallel-plate capacitor, C_{pp}, is given by:

$$C_{pp} = \varepsilon_0 A/d \tag{2}$$

and the magnitude of the capacitive impedance, Z_{pp}, is given by

$$Z_{pp} = 1/2\pi f C_{pp} \tag{3}$$

where f is the frequency of the applied ac signal and π is a constant.

Some dc (or faradaic) current does, however, manage to leak across the double layer due to electrochemical reactions (**1**) and (**2**) taking place at the interface. These reactions experience a charge transfer resistance, R_{CT}, which can be thought of as shunting the nonfaradaic, double-layer capacitance and whose expression can be derived from the Butler–Volmer equation.

For small applied signal amplitudes (7),

$$R_{CT} = \frac{RT}{nF}\frac{1}{i_0} \tag{4}$$

A good electrode, from an electrical point of view, will have a very low value of R_{CT}. Charge will be transferred across the interface almost unimpeded and little voltage will be dropped across the interface. One should note that R_{CT} is inversely proportional to i_0. i_0 is the exchange current [i.e., the current flowing across the interface (in both directions) under equilibrium conditions (no net current flow)]. Simplistically, if an interface can cope with large currents under equilibrium conditions, it will be able to cope well with currents under nonequilibrium conditions. A good electrode system will therefore be characterized by a large value of exchange current or a low value of R_{CT}.

The interface impedance should theoretically be well represented by an equivalent circuit model comprising the double-layer capacitance in parallel with the charge transfer resistance, R_{CT}. Both are in series with R_{TOTAL}, the relatively small resistance due to the sum of the lead and electrolyte resistances.

Complex Impedance Plot. If, for each frequency of ac signal used to measure the impedance, the real part of the measured impedance (Z' or R_S) is plotted on the x axis and the imaginary part (Z'' or X_S) on the y axis of a graph, one obtains a Nyquist or complex impedance plot. The impedance locus for the above simple equivalent circuit model (Fig. 3.) of the interface impedance is plotted on a complex impedance plot in Fig. 4. *Note*: Electrochemists plot–X_S versus R_S and not X_S versus R_S as electrode (and tissue) impedances tend to be capacitive and thus negative. It is generally found easier to look at the plots with the Z'' axis inverted. Low frequency data are on the right side of the plot and higher frequencies are on the left, which is generally the case for electrode interface data.

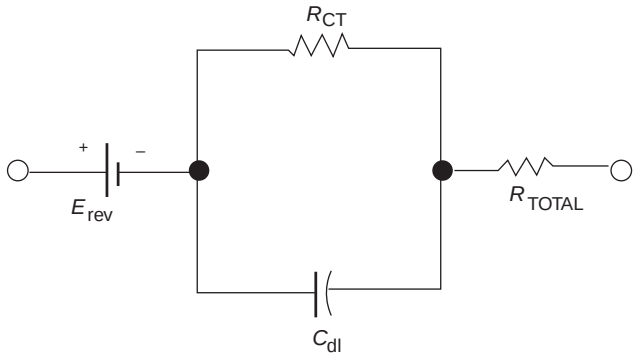


Figure 3. Simple equivalent circuit model of the electrode-electrolyte interface. C_{dl} represents the double-layer capacitance, R_{CT} the charge transfer resistance, R_{TOTAL} the sum of the lead and electrolyte resistances, and E_{rev} represents the reversible or equilibrium potential.

The impedance locus has the form of a semi-circle with high and low frequency intercepts with the real axis at 90° (due to the presence of C_{dl} in parallel with R_{CT}). At very low frequencies, the impedance is equal to $R_{TOTAL} + R_{CT}$, the diameter of the semicircle being equal to R_{CT}. At higher frequencies, the impedance is influenced by the value of the parallel capacitance C_{dl}. As the capacitive impedance decreases with increasing frequency, current therefore flows through it and the total impedance of the parallel combination decreases. The reactive component and the phase angle increases from zero, reaches a maximum value (which depends on the relative sizes of R_{TOTAL} and R_{CT}), and then decreases again toward zero (see Figs. 4 and 5). The frequency at which the reactive component reaches its maximum value (ω_0) is given by $\omega_0 = 1/R_{CT}C_{dl}$ (Fig. 4).

At high frequencies, the impedance is determined by the series resistance R_{TOTAL}.

Bode Plot. Another popular method of presenting impedance data is the Bode plot. The impedance is plotted

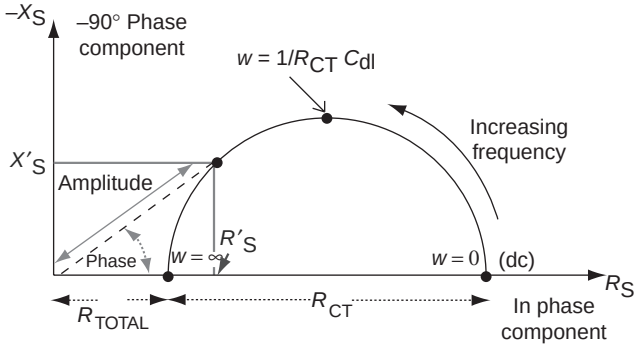


Figure 4. Impedance plot for simple equivalent circuit model of the electrode–electrolyte interface. The impedance locus is semicircular as a result of the parallel combination of C_{dl} (the double layer capacitance) and R_{CT} (the charge transfer resistance), both of which are in series with R_{TOTAL}, the sum of the lead and electrolyte resistances.

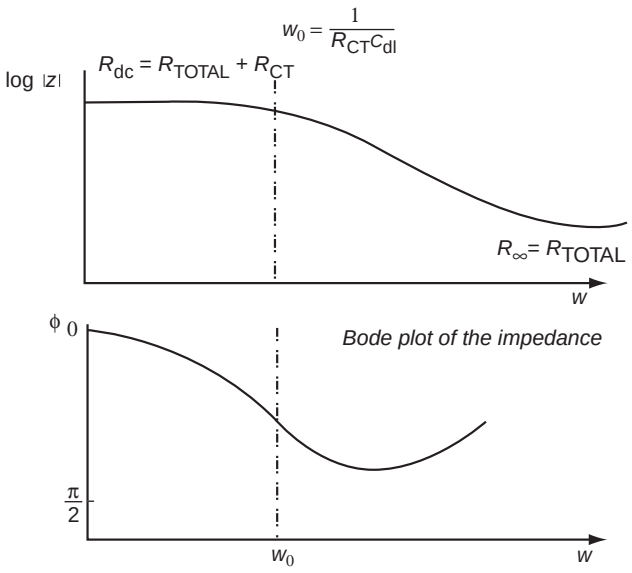


Figure 5. Bode plot for the simple '3-component' equivalent circuit model. The magnitude of the impedance decreases from its low frequency value ($R_{TOTAL} + R_{CT}$) to R_{TOTAL} at very high frequencies. The phase angle of the overall interface impedance increases from 0° at low frequencies, reaches a maximum, which depends on the relative sizes of R_{TOTAL} and R_{CT} and then decreases again.

with log frequency on the x axis and both the absolute value of the impedance ($|Z| = [Z'^2 + Z''^2]^{1/2}$) and the phase-shift ($\phi = \tan^{-1}[Z''/Z']$) on the y axis. Unlike the complex impedance plot, the Bode plot explicitly shows frequency information.

The Bode plot for the electric circuit of Fig. 3 is shown in Fig. 5.

As in the complex impedance plot, the magnitude of the impedance is equal to $R_{TOTAL} + R_{CT}$ at very low frequencies (R_{dc}). The phase angle is zero at this point as the impedance is purely resistive. At very high frequencies, the magnitude of the impedance (R_∞) is equal to that of the series resistance R_{TOTAL}. At frequencies in between these two limits, the interface impedance is influenced by the value of the parallel capacitance C_{DL}. As the capacitive impedance decreases with increasing frequency, current therefore flows through it and the total impedance of the parallel combination decreases. The phase angle increases from zero, reaches a maximum value (generally less than 90° or π/2 rad), and then decreases again toward zero (see Fig. 5).

The above model can be used to explain most key aspects of the electrode–electrolyte interface. It must be pointed out, however, that the equivalent circuit is a gross approximation.

For example, diffusion of ions to the interface from the bulk of the electrolyte (gel or patient) takes place at a finite rate and thus gives rise to impedance to current flow, especially at low frequencies. The diffusion (often termed Warburg) impedance is generally located in series with the charge transfer resistance, both of these being in parallel with the double-layer capacitance. The diffusion impedance has been ignored in the above model as it tends not to be observed for many biomedical electrode systems over the range of frequencies typically used.

A further simplification is the use of a simple capacitance in the above model. Such ideal capacitive behavior is rarely observed with solid metal electrodes. Instead, an empirical pseudo capacitance or constant phase angle impedance, Z_{CPA}, is often used that has a constant phase angle, much like a capacitor.

$$Z_{CPA} = K(i\omega)^{-\beta} \tag{5}$$

where K is a measure of the magnitude of Z_{CPA} and has units of $\Omega s^{-\beta}$, and β is constant such that $0 < \beta < 1$. The phase angle of this empirical circuit element ($\phi = \beta\pi/2$ radians or $90\beta°$) generally lies between 45° and 90° (9). Typically, β has a value of 0.8 for many biomedical electrode systems.

Fricke (10) used the term polarization to describe the constant phase angle impedance and postulated that it was due to spontaneous depolarization of the electrode. Although he did not enlarge on the hypothesis, many authors have used Fricke's terminology over the intervening years. The present author must concur with Cole and Curtis' observation that "the use of the term polarization for describing the unexplained effects occurring at the metal–electrolyte interface is only an admission of our ignorance" (11).

The two most likely causes of the observed constant phase angle impedance are specific adsorption and surface roughness effects (12). With solid biomedical electrodes, the nonideal behavior is probably due to the surface roughness of the electrodes (13), which is supported by reports that roughing an electrode surface decreases the measured value of phase angle.

It is also naive to think that surface effects will only distort the nonfaradaic impedance and will have no effect on R_{CT} as assumed in the above model. It is more realistic that surface effects will affect the parallel combination of C_{dl} and R_{CT} giving rise to skewed (14,15) or distorted (16) arcs. The simple equivalent circuit used in this presentation is, however, a useful approximation that enables qualitative interpretation of much of the published data.

Polarization. Since the work of Fricke (10), the term polarization has been used to describe just about anything associated with the electrode–electolyte interface—frequency-dependence, nonlinearity, noise, and so on. Polarization has been defined as "the departure of the electrode potential from the reversible value upon the passage of faradaic current" (7).

Under equilibrium conditions, the electrode potential E is equal to its reversible potential E_{rev}. When a dc or faradaic current, i_{dc}, is applied to the electrode interface, it must flow through the resistance R_{CT}, which is in parallel with C_{dl}. From Ohm's law, the voltage dropped across this charge transfer resistance will be equal to i_{dc} multiplied by R_{CT} (Fig. 6). The electrode potential E is now given by:

$$E = E_{rev} + i_{dc}R_{CT} \tag{6}$$

The electrode, therefore, is no longer operating at its equilibrium or reversible value E_{rev}. This change in the

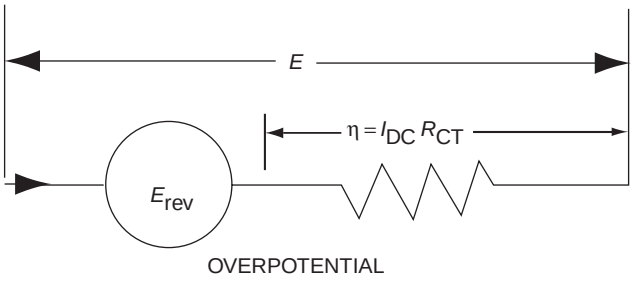


Figure 6. Polarization, the departure of the electrode potential from its reversible value upon the passage of faradaic current.

electrode interface's potential from its equilibrium value is termed polarization. The degree of polarization is measured by the additional voltage dropped across R_{CT}, (or overpotential, η, as it is termed in electrochemistry) where:

$$\eta = /E - E_{rev}/ \tag{7}$$

An ideal nonpolarizable electrode would have a value of R_{CT} equal to zero and, hence, would exist no resistance to faradaic current. The nonfaradaic impedance would effectively be shorted out and the total interface impedance would be zero. In this case, current would pass freely across the interface unimpeded. Measured biosignals, for example, would be unattenuated and undistorted. A perfect electrode system! The electrode potential would always remain constant at its reversible value.

A perfectly polarizable electrode would not permit the flow of any dc or faradaic current as the charge-transfer resistance in this case is infinite. Such an electrode is sometimes termed a blocking electrode. No faradaic charge would cross the interface, even for large overpotentials, and the electrode couples capacitvely with the tissues/electrolyte in this extreme case (Fig. 7).

Real electrodes are, however, neither perfectly polarizable nor perfectly nonpolarizable. Any net current flow across an electrode–electrolyte interface will experience a finite faradaic impedance across which an overpotential will develop.

An electrode system that has a very low value of R_{CT} lets current traverse the interface almost unimpeded, wastes little energy at the interface, has a relatively small overpotential, and has a relatively nonpolarizable electrode system. Such electrode systems are highly sought after, especially when recording small biosignals from the body surface.

Electrodes made of noble metal come closest to behaving as perfectly polarizable electrodes. As these metals are inert, they tend not to react chemically with the surrounding electrolyte or tissue. Noble metals are, therefore, generally used in the construction of implant electrodes where chemical reaction with surrounding tissues must be avoided in order to minimize tissue toxicity problems. Little steady current can pass in such cases as the charge transfer resistance for these electrodes is therefore very large (Fig. 7). The small current that does pass represents the charging and discharging of the double-layer capacitance. A problem, therefore, exists when designing implant electrodes. For a biocompatibility point of view, one requires a noble, hence polarizable, electrode system, whereas from an electrical performance point of view, one requires a nonpolarizable system. A compromise is achieved by using a polarizable electrode and roughening the surface of the electrode, thus decreasing the large interface impedance.

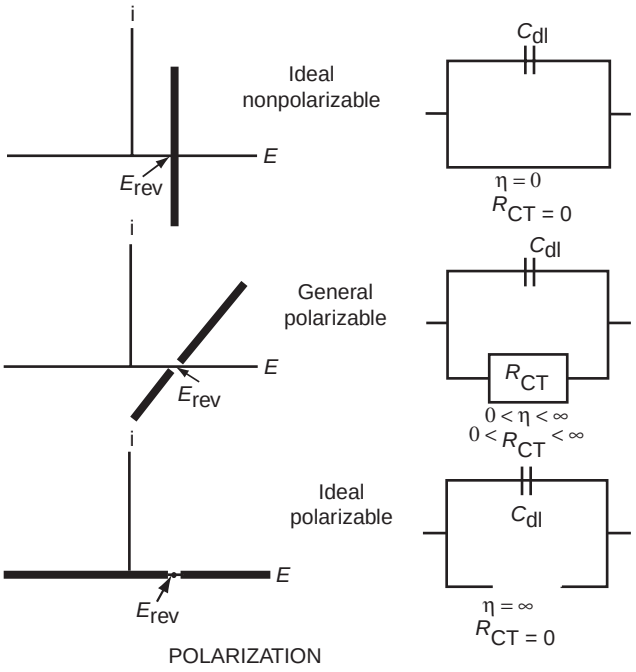


Figure 7. The dc current–voltage plots for Ideal Nonpolarizable, General Polarizable, and Ideal Polarizable electrode interfaces.

Transient Response and Tissue Damage. The response of the electrode system to sine waves of varying frequencies has been considered above (Complex Impedance and Bode plots) as this is a very useful tool in analyzing circuits or, in this case, electrode systems. Equally relevant is the response of an electrode system to voltage and current steps or pulses, as these will approximate therapeutic stimulation applications.

It must be borne in mind that the conversion from electrical to ionic current takes place at the electrode-tissue interface. Based on the simple equivalent circuit model, current can flow either through the parallel resistance or through the double-layer capacitance.

Current flowing through the parallel resistance involves faradaic charge transfer reactions. At the anode, the electrolysis of water and the oxidation of organic compounds can occur. The oxidation of the electrode itself can also occur, which results in the dissolution of metal. At the cathode, hydrogen ions are reduced to form hydrogen gas, which results in a change in pH near the electrode. The new chemical by products in all of these reactions may lead to tissue damage and, hence, faradaic charge transfer reactions must be avoided (17,18). Current must not, therefore, be allowed to flow through R_{CT}.

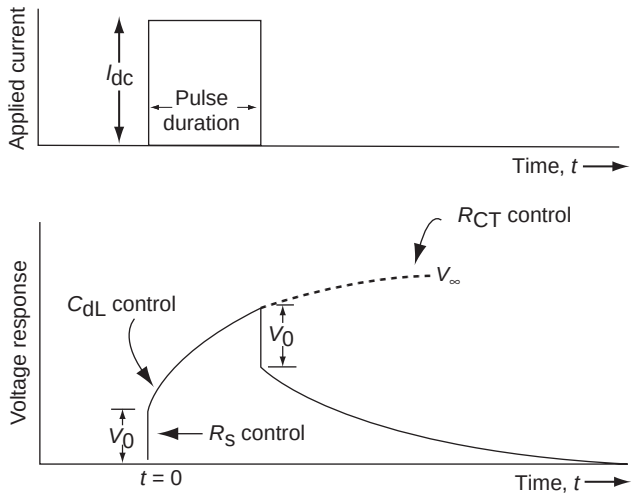


Figure 8. Voltage response to a pulse or step in current.

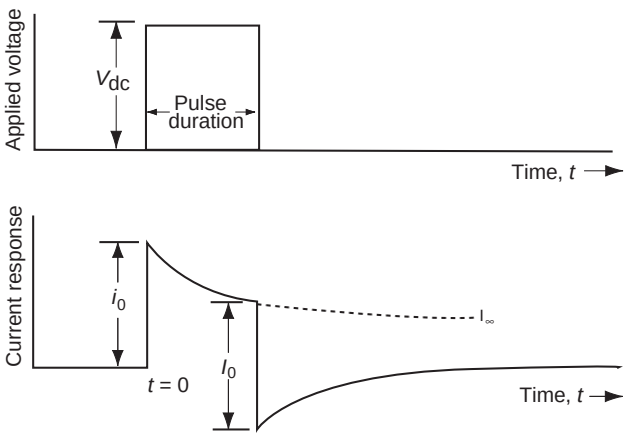


Figure 9. Current response to a pulse or step in voltage.

For current flowing through the double-layer capacitance, no charge actually crosses the electrode-tissue interface. Instead, ions in the tissue are attracted or repelled by charges on the electrode, resulting in transient pulses of ionic current. As no net current flows through the interface and electrochemical reactions are not involved, capacitive current is relatively safe. One must therefore seek, as far as possible, to couple capacitively with tissue when seeking to stimulate tissue without causing trauma.

If one applies a pulse of current of amplitude I_{dc} at time $t = 0$, the voltage response of the electrode-interface equivalent circuit model and, it is believed, the electrode-patient system is as shown in Fig. 8.

$$V(t) = I_{dc}R_{TOTAL} + I_{dc}R_{CT}(1 - \exp[-t/R_{CT}C_{dl}]) \quad (8)$$

At $t = 0$, the applied current flows unopposed through the capacitor and, hence, only sees the series resistance R_{TOTAL}. The initial voltage response is, therefore,

$$V_0 = I_{dc}R_{TOTAL} \quad (9)$$

The voltage response is then observed to gradually increase from V_0. The initial increase in voltage with time is inversely proportional to the magnitude of the capacitance.

For long pulse durations, all of the current will flow through the resistances R_{CT} and R_{TOTAL}. The total resistance seen by the current is, therefore,

$$Z_{(t=\infty)} = R_{TOTAL+}R_{CT} \quad (10)$$

and the limit voltage V_∞ is given by

$$V_\infty = I_{dc}(R_{TOTAL} + R_{CT}) \quad (11)$$

The voltage response will reach this limit value V_∞ in a time period of approximately five time constants, T, where $T = C_{dl}R_{CT}$.

If a perfect step in voltage, V_{dc}, is applied to the electrode system or the three-component model, the current response is as shown in Fig. 9 and given by the equation

$$I(t) = V_{dc}\left\{\left(\frac{1}{R_{TOTAL} + R_{CT}}\right)\right\} + \left(\frac{R_{CL}/R_{TOTAL}}{R_{CT} + R_{TOTAL}}\right)\exp\left[-\frac{R_{CT} + R_{TOTAL}}{R_{CT}R_{TOTAL}C_{dl}}t\right] \quad (12)$$

At the beginning of the voltage step, the resultant current jumps to a relatively large value I_0, where

$$I_0 = V_{dc}/R_{TOTAL} \quad (13)$$

As time passes, the resultant current decreases exponentially with an initial slope inversely proportional to the capacitance C_{dl}. Eventually, the current will reach a limiting value of I_∞, where

$$I_\infty = \frac{V_{dc}}{R_{TOTAL} + R_{CT}} \quad (14)$$

Tissue Damage. The rise in voltage response or the decrease in current response is often attributed in the literature to a mysterious phenomena called polarization. It is, in fact, nothing more than the transient response of a simple three-component circuit model.

When one applies pulses to an electrode-tissue interface (either for implanted or surface stimulation), the applied or resultant current initially flows into the patient via the double-layer capacitance. No reactions are associated with the capacitive flow of current and, hence, few undesirable effects exist when using short duration pulses (i.e. with a duration less than one time constant). As the pulse duration is increased, however, progressively more current flows through the parallel charge transfer resistance and the patient is more at risk due to the reaction byproducts, which is especially true if the pulse is applied for a length of time longer than five time constants giving the voltage or current response time to level off and reach its steady-state value of V_∞ (or I_∞). At this point, most of the current is flowing through the charge-transfer resistance and charge is therefore injected into the tissues via a faradaic process. The byproducts of the electrochemical reactions involved in

the charge-transfer process will diffuse into the skin or tissue, causing chemical injury (19). As the leveling off of the voltage response is indicative of pure charge transfer control, this feature must be avoided at all costs by either avoiding long duration pulses or by using electrode systems with very long time constants, $T = C_{dl}R_{CT}$. In both cases, charge will be largely applied to the body via relatively safe capacitive processes.

It may not always be possible to use sufficiently short pulse durations to avoid tissue damage and still achieve therapeutic effect. It must also be noted that even when using short pulses where the current or voltage response has not leveled off, some current still flows through R_{CT} and a faradaic charge-transfer reaction takes place, with the associated, albeit reduced, problems (18). Over extended periods of stimulation, the byproducts will accumulate in the tissues. In the past, however, the applied waveforms found experimentally to minimize electrode and tissue damage are consistent with the basic goal of minimizing the flow of faradaic current across the electrode interface (20) and thus ensuring little, if any, net transfer of charge. Decreasing the duration, amplitude, and rate of current pulses have all been suggested as each ensures C_{dl} control of the transients and thus avoids, to some extent at least, the undesirable faradaic processes.

An alternative to using short pulse durations is to use electrode systems with long time constants [i.e., with a large value of R_{CT} (faradaic charge-transfer resistance) or C_{dl} (double-layer capacitance)]. As we have seen, noble metals react with difficulty with surrounding tissues and thus have large values of R_{CT} (19). For a given pulse duration, they tend to couple capacitively with tissue. The reactions on both anode and cathode are reversible, involving surface oxygen, and this reversibility explains the observed identical nature of anodic and cathodic waveforms (19). As a result, noble electrodes are widely used for physiological stimulation. Unfortunately, noble metals give rise to large interface impedances. From a biocompatibility point of view, one requires a noble (hence, large impedance) electrode system, whereas from an electrical performance point of view, one requires a low impedance system. A compromise is achieved by using such polarizable electrode materials and roughening the surface of the electrode, thus decreasing the large interface impedance. Roughening the electrode surface gives rise to a significant increase in the interface capacitance, thus increasing further the time constant ($T\uparrow = C_{dl}\uparrow R_{CT}$) and ensuring that the voltage or current response is even more dominated by the highly desirable capacitive processes while decreasing the interface impedance.

It would be a gross oversimplification to attempt to demystify biomedical electrode design by stating that all high performance biomedical electrodes simply have rough surfaces. However, a significant element of truth exists in the statement. For example, terms like activated, sintered, and porous have been used to describe implant electrodes for cardiac pacing and indicate that the electrode fabrication process results, deliberately or otherwise, in a rough-surfaced electrode. It could be argued that it is often the surface finish rather than the electrode metal that gives rise to the favorable electrical properties reported, especially the low interface impedances.

Most electrical stimulation electrodes rely, to some extent, on faradaic mechanisms at the interface between the metal and the tissue. Even in the case of noble metal electrodes with short pulse durations, byproducts of the electrochemical reactions involved will accumulate over time in the tissues when a signal is applied in one direction (monophasic) and will eventually give rise to irritation. With surface stimulation, for example, early designs of electrodes incorporated thick pads of electrolyte-impregnated lint in order to distance the patient's skin from the electrode–electrolyte interface and the undesirable byproducts. Obviously, with implanted electrodes, that short term solution is not possible.

In particular, monophasic anodic pulses must be avoided as they will cause corrosion problems. Additionally, for most applications, cathodic stimulation has a lower threshold than anodic stimulation. Even in the case of monophasic cathodic pulses, however, current flows in only one direction and the chemical reactions at the interface are not reversed.

Biphasic waveforms are preferred in most electrotherapies as the byproducts of the forward reaction are thought to be recaptured by the reverse reaction (21). In using charge balanced waveforms, it is often believed that because no net charge transfer exists across the electrode–skin interface, no net flow of potentially harmful byproducts into the skin. Unfortunately, the electrochemical reaction that occurs to enable the flow of current during the first phase is not necessarily that involved in the second. Byproducts of the first reaction are therefore not always recaptured and may escape from the interface into the patient (22). However, it must be pointed out that the use of charge balanced biphasic waveforms does indeed greatly minimize the problem and, hence, its widespread use in a range of surface and implant applications. Additionally, surface biphasic stimulation is found to be more comfortable than monophasic.

Limit Voltages and Currents of Linearity. Although the non-linearity of the skin's electrical properties has been investigated under a range of conditions, the phenomenon is still far from well understood. "There appears to be, so far, no model available which accounts for both the linear and nonlinear behavior of the electrodes in the frequency and time domains" (23). It is an important feature of an electrode as 'it appears . . . that electrodes often introduce nonlinear characteristics that are erroneously ascribed to the biological system under study (24).

Schwan proposed empirical relationships for the limit current of linearity and the limit voltage of linearity.

Limit Current of Linearity. If has been observed that electrode–tissue interface impedance nonlinear behavior is first evidenced at low frequencies. As the applied current amplitude is increased, progressively higher frequency points are affected. Schwan proposed a limit current of

linearity i_L. He observed that the relationship between the angular velocity of a given impedance point and the current amplitude required to drive it into nonlinearity (deviate by more than 10% from its linear, small-signal value) was well expressed by the empirical relationship

$$i_L = B\,\omega^{\beta} \tag{15}$$

where B is a constant particular to the electrode system and β is the fractional power that appears in equation 5.

Schwan and others (25–28) have observed that this empirical relationship is valid for many electrode systems over wide frequency ranges.

The presence of β (a parameter describing the frequency dependence of the linear interface impedance) in a relationship describing the nonlinearity of the system was found most intriguing.

The solution to this mystery is quite simple when it is approached from the right direction. Generally, researchers have assumed that the observed nonlinear behavior is attributable to the high frequency Z_{CPA} impedance (Eq. 5), which they observe under linear, small-signal conditions and over the limited frequency ranges they use. However, in parallel with Z_{CPA} is the charge transfer resistance R_{CT}, which, in the linear range, has a very large value R_{CT}^{0}, where

$$R^{0}_{CT(0)} = \frac{RT}{nF}\,\frac{1}{i_0} \tag{16}$$

As a result, its contribution is either not observed or ignored.

The value of the charge-transfer resistance can be derived from the Butler–Volmer equation and is very nonlinear, decreasing rapidly with applied signal (ac or dc) amplitude. Compared with R_{CT}, Z_{CPA} is relatively linear. R_{CT} is therefore the source of the observed nonlinear behavior.

As the applied current amplitude is increased, the charge transfer resistance decreases rapidly, causing the diameter of the impedance locus to decrease. As the low frequency end of the arc is dominated by the charge transfer resistance, the effects of such nonlinearity will be first evidenced at these frequencies. Low frequency points are therefore the first to deviate significantly (by more than 10%) from their small-signal, linear values, as observed by Schwan and others. As the applied signal amplitude is further increased, the diameter of the impedance locus decreases further, and progressively higher frequencies are affected (29,30).

Simplistically, it can be shown that the following approximations can be made over limited ranges of frequency or applied signal amplitude:

- Approximate relationship between applied current and R_{CT}

$$i \propto 1/R_{CT} \tag{17}$$

- Approximate relationship between R_{CT} and the frequency at which nonlinearity occurs

$$R_{CT} \propto \omega^{-\beta} \tag{18}$$

Then, by cancelling R_{CT}, in the above two equations,

- Approximate relationship between applied current and the frequency at which nonlinearity occurs

$$i_L \propto \omega^{\beta} \tag{19}$$

as found by Schwan. The presence of β in the expression of an electrode system's nonlinear behavior is therefore simply due to the presence of a very nonlinear resistance in parallel with a relatively linear, frequency-dependent Z_{CPA}. A more accurate calculation based on the equivalent circuit model outlined above and the Butler–Volmer equation was published (29).

Limit Voltage of Linearity. Schwan and others (25,28) also postulated that the electrode–electrolyte interface impedance becomes nonlinear at a certain limit voltage, V_L, which they found to be independent of the frequency of the applied signal.

Using the Butler–Volmer equation and the equation for the impedance of the equivalent circuit model, the voltage limit of linearity can be calculated for a range of frequencies (31). It can be shown that the charge-transfer resistance decreases pseudo exponentially with applied voltage amplitude, initially causing low frequency impedance points on the locus to deviate from their small-signal values (32,33).

At very low frequencies, such that $\omega \rightarrow 0$, the voltage limit of linearity, V_L, approximates to the voltage at which the charge-transfer resistance decreases by 10% from its small-signal value, which occurs at $V_L = 40/n$ mV, where n is the number of electrons per molecule oxidized or reduced (31).

As the applied voltage is increased above this low frequency limiting value, the charge transfer resistance further decreases and affects progressively higher frequency points (i.e., become nonlinear). The derived log (f) versus V_L plot is found to be a straight line over a wide range of frequencies. V_L is observed to increase only very slightly with frequency, which would agree qualitatively with Onaral and Schwan's results (28), where V_L increased from 106 to only 129 mV over the frequency range of 10 mHz–100 Hz for platinum electrodes in saline, which would also explain why, in the past, V_L has been assumed constant and independent of the applied frequency.

Electrode Metals. As biocompatibility is of great importance in implants, implant electrode materials are generally confined to those that are essentially inert and do not react with the surrounding tissues. As cardiac pacing electrodes were among the first implanted and have had a long, generally successfully and well-researched history, most conclusions drawn on the suitability of materials for implant electrodes are based on pacing electrodes.

Implant electrodes are and have been generally made from noble metals such as gold, platinum, iridium, rhodium, and palladium. Platinum has been the most widely used as it has excellent corrosion resistance and produces relatively low polarization (34). Platinum, however, is mechanically relatively soft and for many applications is

alloyed with much harder iridium, producing platinum-iridium. Other noble metal alloys that have been used include gold–platinum–rhodium, platinum–rhodium, and gold–palladium–rhodium.

Passive metals, such as titanium, tantalum, zirconium, tungsten, and chromium, have been successfully implanted. Titanium has been widely used because it forms a nonconducting oxide layer at the surface. This coating prevents charge transfer at the electrode interface. Titanium, therefore, exhibits a high resistance to corrosion. Stainless steel is similar in that it acquires a protective oxide layer that renders it inert. Although stainless steels were used in early pacing electrodes, they do not appear to have the required corrosion resistance for long-term use. Stainless-steel pacing electrodes were discontinued after the 1960s because of unreliable corrosion resistance (34,35).

Some early pacing electrodes were made of Elgiloy (an alloy of Fe, Ni, CO, Cr, and MO from Elgin Watch Co.) However, Elgiloy has marginal corrosion resistance and produces a relatively high polarization overvoltage. It was discontinued in the 1980s. Carbon is an inert, nonmetallic element that has similar electrochemical characteristics to noble metals and continues to be used successfully as an implant electrode. Materials such as zinc, copper, mercury, nickel, lead, copper, silver, silver chloride, iron, and mild steel have been found toxic to body tissues and are normally not used.

Biocompatibility has been defined as the ability of a material to perform with an appropriate host response in a specific application (36). Strictly speaking, no such thing as a biocompatible material exists as an implant's biocompatibility will also depend on a range of variables including its shape and surface finish.

Stimulation threshold is a key parameter in implant stimulation electrode design. When activitated vitreous carbon electrodes were first introduced in pacing electrodes, they were found to have relatively low chronic thresholds. These thresholds were thought to be the result of the superior biocompatibility of the carbon electrode. Other researchers similarly interpreted the low thresholds observed for their new exotic materials such as indium oxide, titanium nitride, and semimetal ceramics. Stokes (34), however, concluded that "material selection appears to have little or nothing to do with threshold evolution—as long as the material is biocompatible and reasonably corrosion resistant. Thus our experiments with biocompatible materials such as carbon, titanium, platinum, iridium oxide, and many more have all produced about the same results when tested as polished electrodes, all other factors held equal". Stokes went on to point out "while the bulk properties of an electrode material are important, it is the electrode-tissue interface that determines the electrode's performance. In fact, the surface microstructure of the electrode is critical" (34). It would appear that the microstructure of an electrode surface may affect cellular adhesion and activation, thus reducing the foreign body response. It is, therefore, the surface structure of many of the new materials (resulting from their fabrication process) that gives rise to the observed positive effect on threshold evolution over time, rather than the biocompatibility of the bulk material.

Another advantage of porous and microporous implant surfaces is their reduced interface impedance. Although interface impedance is generally less critical for implanted stimulation electrodes, many such electrodes (e.g., implanted pacing electrodes) are used to monitor biosignals as well as to deliver the required stimulation impulses. Decreased interface impedance helps in this regard.

Implanted biosignal monitoring electrodes require stable potentials as well as low interface impedances to minimize biosignal recording problems. These metals have high positive standard electrode potentials (E^0 in Eq. 1) and are the lowest ones on the electromotive series. As noble metal electrodes do not tend to react chemically with the electrolyte, the Nernst equation is not defined and the measured potential is often influenced more by any traces of impurities on the surface than by the intrinsic properties of the metal itself. The electrode potential can drift randomly, especially immediately following implantation. It may fluctuate widely under apparently identical circumstances, which is an inherent disadvantage of noble materials.

External biosignal monitoring electrodes can generally use high electrical performance nonnoble materials such as silver–silver chloride without fear of biocompatibility problems (2). Silver–silver chloride has been found to be an excellent electrode sensor material as, when it is in contact with a chloride gel, it has the following characteristics:

1. A low, stable electrode potential.
2. A low level of intrinsic noise.
3. A small value of charge transfer resistance (i.e., it is relatively nonpolarizable).
4. A small interface impedance.

A silver–silver chloride electrode is generally made by the deposition of a layer of silver chloride onto a silver electrode. Silver chloride is a sparingly soluble salt and, thus, effectively provides the silver electrode with a saturated silver–chloride buffer, which facilitates exchanges of charge between the silver electrode and the sodium chloride environment of the gel and human body. The system behaves as a reversible chloride ion electrode, and the Nernst potential, in this case, depends on the activity (which is closely related to concentration) of the environment chloride ions and not on that of the silver ions. The potential of this electrode is, therefore, quite stable (as well as small) when the electrode is placed in an electrolyte containing Cl as the principal anion—as is the case in the human body and electrode gels (2).

Electrical noise (potential fluctuations) can occur spontaneously at the electrode interface without any physiological input. Ag/AgCl electrodes have been shown to be particularly stable and resistant to noise (37).

A silver–silver chloride electrode has a relatively large value of exchange current density (2) (Eq. 4) and, hence, a very low value of charge transfer resistance, R_{CT}. Charge is transferred across the interface with relative ease and little voltage is dropped across the interface. The electrode therefore operates close to its equilibrium or reversible

potential. Ag-AgCl electrodes are, therefore, relatively nonpolarizable.

When a smooth-surfaced electrode is chlorided, the AgCl deposit can give rise to a very rough surface and thus to relatively very low interface impedances (37,38). K, the magnitude of the interface pseudocapacitance (Eq. 5), is observed to decrease following the deposition of an AgCl layer (39). However, although AgCl facilitates the interfacial electrochemistry, it is very resistive having a resistivity of around 10^5–$10^6\,\Omega\cdot$cm (2). As the layer thickness increases, the series resistance, R_{TOTAL} will therefore increase. This series resistance dominates the very high frequency interface impedance, and the latter will also increase with chloride deposit. Therefore, an optimal layer thickness exists, for a given frequency, that decreases the interface impedance and yet does not significantly increase the series resistance, R_{TOTAL} (29). The optimal silver chloride layer thickness consequently depends on the frequency range of interest (40).

Tin-stannous chloride, a material somewhat similar to silver–silver chloride, was used in some biosignal electrodes (41).

Electrode material and high electrical performance is generally less critical for external stimulation electrodes such as TENS and external pacing electrodes (current density distribution is the key concern). The majority of commercially available TENS electrodes are molded from an elastomer such as silicone rubber or a plastic such as ethylene vinyl acetate and loaded with electrically conductive carbon black. Mannheimer and Lampe (42) pointed out that the only tangible disadvantage with having a large electrode interface impedance is that more power will be required from the stimulator to drive the stimulating current through the electrodes into the patient.

Graphite-loaded polyesters and similar materials are used in external pacing electrodes, for example. Some are constructed using tin as the metal layer. In early electrodes, the combination of tin and the chloride-based gel gave rise to pitting of the metal. Improvements made to the gels and the use of high purity tin have effectively removed this problem.

Although silver-silver chloride has been and still is used in some external electrostimulation electrodes, it should be used with care. Silver chloride is deposited electrolytically and can therefore be either removed by the passage of current or a thicker, high resistance layer deposited, depending on the polarity of the electrode, which can be a significant problem in iontophoretic transdermal drug delivery and may cause problems in multifunction pads, which include a silver–silver chloride layer to enable distortion-free monitoring of the ECG through electrodes designed to deliver the pacing or defibrillation impulses.

The Skin

Structure of the Skin. The skin is a multi layered organ that covers and protects the body. It is made up of three principal layers—the epidermis, the dermis, and the subcutaneous layer. (*Note:* In the literature, variations exist in the terminology used to denote these layers.)

The epidermis, the outermost layer, is around 100 μm thick, depending on body site. It is the strongest layer, providing a protective barrier against the outside hostile environment. Unlike any other organ of the body, the epidermis renews itself continually. It can be subdivided into several layers, with the basal layer forming the innermost layer and the stratum corneum the outermost layer. Cells in the basal layer constantly multiply and, as they are pushed up toward the skin's external surface, the cells undergo changes. Eventually, layers of compacted, flattened, nonnucleated, dehydrated cells (called corneocytes) form the stratum corneum. These dead cells are continuously being shed and are replaced from the underlying epidermal layers. The intercellular spaces between corneocytes are occupied by arrays of bilaminar membranes with the morphological features of polar lipids (43). This matrix appears to serve to bind the cells and the stratum corneum has been described in terms of corneocyte bricks surrounded by lipid mortar (44). On average, the stratum corneum comprises around 20 cell layers thick and has a thickness of around 10–15 μm. Thickness will, however, vary with the number of cell layers making up the stratum corneum and the state of hydration. On some body areas, it can be several hundred micrometers thick. The epidermal layer is traversed by numerous skin appendages such as hair follicles, sebaceous glands, and sweat glands.

The underlying layers of the epidermis are, in contrast, a relatively aqueous environment. The transition from an essentially nonconductive, lipophilic membrane (the stratum corneum) to an aqueous tissue (viable epidermis and dermis) gives rise to the skin's barrier properties.

The dermis is the second layer of the skin and, with an approximate thickness of 2 mm, is considerably thicker than the epidermis. It is formed from a dense network of connective tissue made of collagen fibers, giving the skin much of its elasticity and strength. Embedded in the dermis are blood vessels, hair follicles, sebaceous and sweat glands, and several types of sensory nerve endings.

The final layer of the skin (the subcutaneus layer) is found beneath the dermis layer. It contains structures of connective tissues and enables the skin on most parts of the body to move freely across the underlying bone structures. It is one of the body's areas for fat storage and acts as a cushion to protect delicate organs lying beneath the skin.

Skin Impedance

Electrical Properties of the Skin. As the stratum corneum is relatively nonconductive, it presents a high impedance to the transmission of electric currents. As a result, the impedance of the skin is the largest component of the overall interelectrode impedance (Fig. 10). Nonetheless, due to the stratum corneum's dielectric properties and its thinness, it permits capacitive coupling between a conductive metal electrode placed on the skin surface and the underlying conductive tissues. One can imagine the relatively nonconductive stratum corneum sandwiched between the conductive electrode and the conductive tissues underlying the stratum corneum forming a parallel-plate capacitor. The stratum corneum's electrical impedance is, therefore, often represented by a simple capacitor, C_{SP}. (The subscript SP refers to Skin and Parallel.)

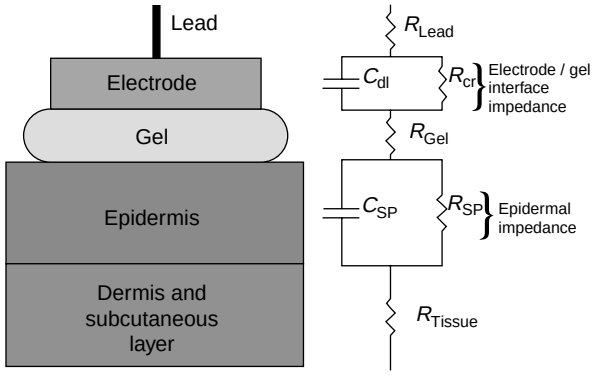


Figure 10. Schematic representation of the skin and its equivalent circuit model.

Some ions do, however, manage to cross the stratum corneum via paracellular pathways and through the skin's appendages (hair follicles, sweat ducts, sebaceous glands, imperfections in the integrity of the skin). As skin appendages extend through the stratum corneum barrier they can act as shunts to the interior. The flow of ionic current can be represented electrically by a large resistance, R_{SP}, in parallel with C_{SP}.

The underlying layers of the epidermis are, in contrast, relatively conductive and can be collectively represented by a tissue resistance, R_{Tissue}.

A simple equivalent circuit model of the overall electrode-gel-skin system is therefore shown in Fig. 10 and includes the electrode lead resistance, R_{Lead}; the electrode–gel interface impedance (the double-layer capacitance, C_{dl}, in parallel with the charge transfer resistance, R_{CT}); the gel resistance, R_{Gel}; the skin impedance (the parallel combination of a capacitance, C_{SP} and a resistance, R_{SP}), and the underlying tissue resistance, R_{Tissue}.

It must be borne in mind that this equivalent circuit model is a simplification of the rather complex electrical properties of the skin. For example, it has been found experimentally (45,46) that the capacitance of the skin is better described by an empirical, constant phase angle impedance, Z_{CPA}, where

$$Z_{CPA(S)} = K_S(j\omega)^{-\alpha} \qquad (20)$$

[which is similar to the empirical expression for the pseudo capacitance often used to represent the nonideal capacitive properties of the electrode interface's double-layer capacitance (Eq. 5)].

K has units of $\Omega \cdot s^{-\alpha}$. The parameter α is constant such that $0 \leq \alpha \leq 1$. The fractional power, α, of the capacitive impedance has been found to be related to the degree of hydration of the stratum corneum (47). If the epidermal layer behaved as a simple capacitance, α would equal unity. The actual value of α, normally around 0.8–0.9, is a measure of the deviation from this ideal behavior.

The use of C_{SP} will, however, be sufficient for this present review. The lead resistance, the gel resistance, the tissue resistance, and the electrode–gel interface impedance are all relatively small in comparison with the large skin impedance. Skin impedance, therefore, generally dominates and will be studied in more depth.

The Skin's Parallel Capacitance, C_{SP}. It was suggested above that the electrode–skin interface can be approximated by a capacitor with the stratum corneum forming the dielectric layer sandwiched between the electrode and the underlying tissues that form the conductive plates. It can be seen from Eq. 2 that the skin's capacitance will increase as the thickness of the stratum corneum decreases, its dielectric constant increases, or the area of the electrode increases.

The number of cell layers in the stratum corneum can range from 12 to 30 (48). Epidermal thickness can, therefore vary greatly for different body sites within the range of about 10 to well over 100 μm (49). The stratum corneum can be, for example, as thick as 400–600 μm in the palm and plantar areas and as little as 10–20 μm on the back, legs, and abdomen (50). The value of the capacitance of the skin is related to the thickness and composition of the stratum corneum and has a typical value in the range $0.02–0.06\ \mu F \cdot cm^{-2}$ when measured using electrodes with "wet" electrolyte gels several minutes following electrode application (51,52). As the stratum corneum is typically at least 10 times as thick on the palms of the hands and soles of the feet as compared with other body areas, the skin capacitance at these points is considerably smaller than at other sites on the body. The stratum corneum on the face and scalp is not as thick as on other body parts and is characterized by large capacitance values.

Dark-skinned subjects have stratum corneum layers that are more dense and contain more layers of cells than fair-skinned subjects (48). Not surprisingly, they are characterized by skin capacitances that are much lower (skin impedances, Eq. 3, much higher) than those for fair-skinned subjects. One should therefore take care when assessing a new electrode system or associated device that they are tested on a range of subjects and skin sites. What may work well on a subject with low skin impedance in a warm and humid environment may be found later to fail on a high impedance subject, especially in a cold or dry environment.

The Skin's Parallel Resistance, R_{SP}. Although the stratum corneum does not easily allow foreign substances to traverse it, some current, carried by ions, manages to flow through it. The difficulty or resistance, this current experiences in passing through the skin is represented in the equivalent circuit (Fig. 10) by the parallel resistance, R_{SP}.

The skin's resistance is highly dependent on the presence and activity of sweat glands and on the presence of other appendageal pathways. An average human skin surface is believed to contain between 200 and 250 sweat ducts on every square centimeter (53). The density of sweat glands varies greatly over the body surface with a value of approximately 370 per cm^2 on the palms of the hands and the soles of the feet and a value of approximately 160 per cm^2 on the forearm (49). The diameter of the ducts can range from 5 to 20 μm. It is, therefore, not surprising that R_{SP} is reported to vary greatly from patient to patient, from body site-to-body site, and with time. The measured values of R_{SP} are much smaller on areas with high densities of sweat glands, such as the palms of the hands (in spite of the thicker stratum corneum layer), especially when the

glands are active in response to thermal or psychophysiological stimuli.

An average human skin surface is reported to contain between 40 and 70 hair follicles per square centimeter (53). The presence of a high density of hair follicles (which act as low resistance shunts) gives rise to a very low value of skin parallel resistance, R_{SP}. However, this observation is counterbalanced by the difficulty in making firm mechanical and electrical contact to hirsute body sites or patients. In such cases, the skin impedance is very large at best. Generally, the electrodes fall off and, hence, require the shaving of the skin site prior to electrode application.

Observed intersite and interpatient variations in skin impedance tend to be due to large variations in R_{SP}. In the low frequency range, dominated by R_{SP}, regional differences in skin impedance were observed by Rothman (54), Lawler et al. (55), and Rosell et al. (56). Low frequency skin impedance was observed to decrease in the following order: thumb, forearm, abdomen and, smallest of all, forehead. Similarily, Almasi and Schmitt (57) observed the low frequency skin (10 Hz) impedance to decrease in the order of outer forearm, leg, inner forearm, back, chest, earlobes, and forehead. The forehead appears to have a very low skin impedance value (58), presumably as a result of the stratum corneum on the face and scalp being thinner than that on other body parts (48) and the presence of a high density of sweat glands. Almasi and Schmitt (57) plotted their average impedance values for the body sites on a complex impedance plot and found that most of the points lay along a "smooth common locus of monotonically increasing phase angle and impedance magnitude." This behavior was successfully interpreted by McAdams and Jossinet (32), who showed that such frequency loci were formed when the skin's parallel resistance varied greatly from site to site while the skin's capacitance remained relatively constant. Two body sites did not fit the locus and, hence, the physical explanation; these sites were the palm and fingertips. These body sites have much larger epidermal thicknesses and, hence, have skin capacitance values much smaller than other body sites.

One must be very careful when assessing different electrode designs or gels. Testing different electrodes on different patients is certain to give misleading results due to the intersubject variations, unless, of course, large numbers of subjects are used and statistically significant differences are observed.

R_{SP} varies greatly over time due to a number of parameters including room temperature and psychophysiological stimuli. The latter effect is exploited in so-called lie detectors. Schmitt and Almasi (59) reported that a considerable daily variation exists in a given subject, and seasonal changes have also been reported (60). Testing a range of electrodes on the same subject but on different days is, therefore, not optimal either, as day-to-day variations in skin impedance, especially fluctuations in R_{SP}, will nullify the validity of this approach. For example, Searle and Kirkup (61) found that the diurnal variations on a given subject for a given electrode was much larger than any difference between the range of electrode designs they tested in any one recording session. It should be further noted that the electrode test sites should be allowed to recover for several days between experiments to enable the skin to recover. For example, pealing off an adhesive electrode will remove some of the underlying stratum corneum. Any electrode subsequently tested on the site will benefit from this prior skin stripping (see below).

Electrode designs must, therefore, be compared *in vivo* by testing them at the same time on the same subject. One must still bear in mind the significant differences in skin impedance that exist over the subject's body, as outlined above. Even testing the electrodes at the same time on a limb of a given subject remains problematic. The different skin sites involved, even if located close together, will give rise to significant differences in the measured electrode–skin impedances, which may be wrongly attributed to the electrode designs or gels under test. Searle and Kirkup (61), for example, showed that testing a range of dry electrode metals on the inner forearm gave rise to potentially very misleading results. Electrodes placed closer to the wrist gave rise to lower impedances due to the presence of a higher concentration of sweat glands.

Electrodes must therefore be repeatedly tested at the same time, under the same conditions, varying their relative positions in order to clearly establish their relative performances. McAdams et al. developed a four-channel impedance monitoring system to enable the simultaneous comparison of electrode designs/gels (62).

Skin Potential Motion Artifact. A potential difference E_S, given by the Nernst equation, exists across the epidermis as a result of ionic concentration differences. This potential varies from patient to patient, from site to site, and depends on gel composition (if used) and skin condition.

The skin surface is normally negative with respect to the inside of the body. Skin potential becomes more negative when sweat glands are active, and palmer and plantar surfaces, with their higher sweat gland concentrations, are the most negative. Increasing gel concentrations of NaCl or KCl also render the site more negative. The parameter E_S has a typical value of 15–30 mV (63).

The dependence of the skin potential on the thickness of the epidermal layer is important to many ECG recording applications. If the thickness of the layer is changed by stretching or pressing down on the skin, the skin potential can vary by 5–10 mV compared with, for example, the 1 or 2 mV ECG signal. As these fluctuations generally result from patient movement, they are termed motion artifact. Motion or skin-deformation artifact is a serious problem during exercise cardiac stress testing of patients on treadmills or exercise bicycles, during ambulatory monitoring, and while monitoring patients lying in bed (64,65) (Fig. 11).

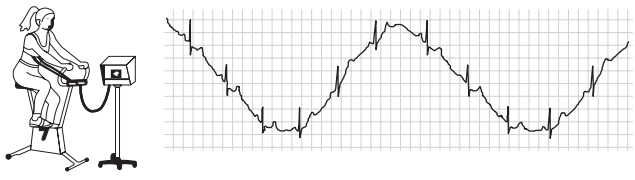


Figure 11. Disturbance of biosignal due to patient movement. (Redrawn from Ref. 65.)

Abrading or puncturing the skin is often used in stress testing to remove or bypass the problem source. Although skin potential increases with gel concentration, artifact gradually decreases with time as the conductive electrode gel soaks into the skin and renders the stratum corneum more conductive. High concentration gels are often used for short-term diagnostic applications where the risk of skin irritation is outweighed by the need for clear traces.

In general, hydrogel-based electrodes (see below) should not be used for stress testing or in other monitoring applications that are likely to suffer motion artifact problems. Hydrogels tend not to hydrate the skin and, hence, do not actively attack the source of the problem. The same comment applies for dry (gel-less) electrodes.

In stress testing and ambulatory event monitoring, modified electrode locations are used to avoid muscular or flabby areas of the body and thus minimize skin-deformation (and EMG) artifact. Stress loops are formed in the connecting leads, which are taped to the patient and used to avoid direct pull on electrodes and the underlying skin. The use of foam-backed electrodes tends to absorb any pull on the electrode and minimizes artifact.

Electrode Gels and Their Effects. Dry electrodes are successfully used in some monitoring applications. Suitably designed gel-less electrodes have advantages when used in the home environment where the patient may not remember or have the time to apply gel electrodes prior to use (66).

For many home-based monitoring applications, electrodes are manufactured from noncorroding materials such as stainless steel, which can be repeatedly washed and reused. Unfortunately, such polarizable materials give rise to poor electrical performances. In order to ensure good, stable electrode potentials, silver–silver chloride electrodes should be used (see below).

Jossinet and McAdams (67) demonstrated that the impedance of a dry electrode decreases pseudoexponentially due to the gradual buildup of sweat under an occlusive, gel-less electrode and the resultant progressive hydration of the underlying skin. Searle and Kirkup (61) reported that the decrease in skin impedance of dry electrodes is polyexponential and requires two time constants, one very short (~45 s) and the other almost 10 times longer (~450 s), possibly indicating two different processes at work.

Given that the surface of the skin is irregular, a flat dry electrode will initially only make contact with a few 'peaks' on the skin surface. Therefore, a smaller effective contact area exists than one would otherwise expect. However, as sweat builds up under the occlusive, dry electrode, a better contact with the skin will result in a relatively rapid increase in the measured value of C_{SP}. Human sweat contains a small amount of sodium chloride [~0.1–0.4% NaCl (49)], and hence serves as a weak electrolyte. It is suggested by the author that this accounts for the shorter time constant. (The longer time constant is probably indicative of the progressive hydration on the underlying skin resulting in a gradual decrease in R_{SP}.) As will be outlined below, R_{SP} is observed to decrease with a time constant of around 10 min in the presence of an electrolyte gel, which agrees quite well with the 7.5 min observed by Searle and Kirkup (61).

Before leaving gel-less electrodes, it should be pointed out that in certain applications that employ very high frequency signals, such as electrical impedance tomography (EIT), the use of a gel pad may not be needed as it will contribute a small but significant contact resistance to the desired measurement (5). In such instances, the use of a very thin spray of moisture onto the electrode surface prior to its firm application to the patient's skin may be all that is required. Profiled dry electrodes firmly pressed onto the skin may also be adequate for certain home-based biosignal applications. If skin impedance is a problem with standard button electrode designs, this can be addressed by increasing the electrode area in the noncritical axis. For example, long, narrow, dry electrodes are used for precordial ECG recording, which enable a large contact area while ensuring sufficient interelectrode distances on the chest (66).

Electrode gels serve (*1*) to ensure a good electrical contact between the electrode and the patient's skin, (*2*) to facilitate the transfer of charge at the electrode-electrolyte interface between the two kinds of charge carrier (electrons in the electrode and ions in the gel), and (*3*) to decrease the large impedance of the stratum corneum.

Two main types of electrode gel exist, viz. wet gels (often described as pastes, creams, or jellies) and hydrogels.

Wet gels are generally composed of water, a thickening agent, a bactericide/fungicide, an ionic salt, and a surfactant (68). The ionic salt is present to achieve the appropriate electrical conductivity of the gel, which will depend on the specific application. As the major portion of ions present in tissue fluids and sweat are sodium, potassium, and chloride (Cl^-), in order to ensure biocompatibility, the ionic salts most commonly used in electrode gels are NaCl (sodium chloride) and KCl (potassium chloride). High concentrations of these salts tend to be better tolerated by the body than other salts. The ions in the gel serve not only to ensure electrical conductivity of the gel but to decrease the skin impedance by diffusing into the skin due to the existing concentration gradient. A relatively high concentration of electrolyte will also decrease the value of the charge transfer resistance (thus rendering the electrode more nonpolarizable).

When a standard pregelled wet electrode is applied to the skin, the gel rapidly fills up the troughs on the electrode and skin surfaces, thus ensuring maximum effective contact area. The skin capacitance, C_{SP}, is therefore observed to initially increase rapidly in value following electrode application and then to remain relatively constant (32). [A similar effect was probably noticed by Searle and Kirkup (61) as a result of sweat accumulation under a dry occlusive electrode.] Although C_{SP} does not exhibit a strong time dependence, it does vary with the electrolyte composition and concentration (49), increasing with increasing concentration (69).

Following electrode application, the skin's parallel resistance, R_P, generally decreases with time in a pseudo exponential manner as the ions in the gel diffuse through the skin rendering it more conductive (32,70) (see Fig. 12).

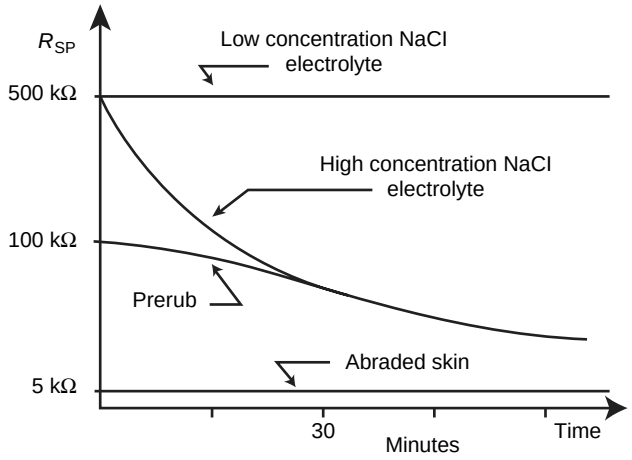


Figure 12. Variation of skin resistance with time for a range of gel concentrations and skin preparation techniques (1).

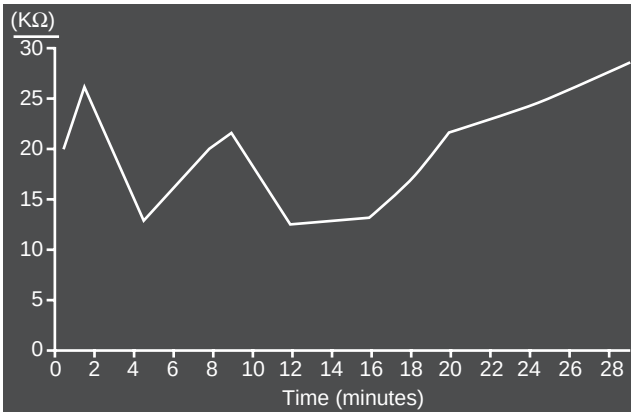


Figure 13. Variation of skin resistance with time for a hydrogel-based electrode. Fluctuations are due to variations in subjects state of relaxation/arousal (32).

It has been observed, however, that when a cold gel is applied to a skin site, the measured value of R_{SP} is observed to initially increase (32,56), which is attributed to the cold gel causing the sweat pores to contract. Once the gel and skin has warmed up, the value of R_{SP} is observed to decrease as the electrolyte ions diffuse through the epidermal layer.

The skin temperature effect should be borne in mind when assessing a range of electrode designs. If, for example, the patient/subject removes his/her shirt just before the tests, the electrodes tested at the start of the experiment will have an advantage (i.e., smaller skin impedance) over those tested later as the uncovered skin sites will gradually cool down with time following removal of the shirt. Meaningful *in vivo* assessment of electrodes is not straightforward, and wrong conclusions can very easily be made by the unaware or the unscrupulous.

The time constant for the skin's parallel resistance, R_{SP}, appears to be inversely proportional to the concentration of the gel. The decay has a time constant of around 10 min (1), thus indicating that it takes almost 1 h for the electrode-skin impedance to decrease to its lowest value. For example, 50/60 Hz interference, linked to mismatch of electrode-skin impedances, is often observed experimentally to decrease with time. One should, therefore, where possible, apply the electrodes to the patient first, for example, before setting up the rest of the measurement system, to enable the skin impedance to decrease as much as possible.

High salt concentrations give rise to a more rapid diffusion of ions into the skin and a more rapid decrease in the skin's parallel resistance, R_{SP} (1,32) (Fig. 13). Such aggressive gels tend to be used in short-term biosignal monitoring applications such as stress testing, where instant, high quality traces are required (71). Biological tissues cannot tolerate long-term exposure to salt concentrations, which depart significantly from physiological levels [~0.9% NaCl for body fluids and around 0.1–0.4% NaCl for human sweat (49)]. Aggressive gels (5% NaCl) should not be used, for example, for the long-term monitoring of bed-ridden patients or for the monitoring of neonates. In the latter case, the incompletely formed skin is very susceptible to skin irritation problems. In any monitoring application, aggressive gels should not be used in combination with skin abrasion (see below). It is especially to be avoided in longer term monitoring applications where the removal of the body's defensive barrier coupled with the long-term exposure to an aggressive gel will lead to sever discomfort to the patient.

The second kind of electrode gel commonly used in electrode systems are hydrogels. Hydrogel-based electrodes have recently become popular for numerous biomedical applications including resting ECG. Hydrogels are solid gels, which originally incorporated natural hydrocolloids (e.g., Karaya gum a polysaccharide obtained from a tree found in India) (68). The use of natural hydrocolloids give rise to variable performances and, in some cases, an unattractive color. Synthetic (e.g., polyvinyl pyrrolidone) hydrocolloids are now widely used.

The use of such solid gels entails numerous advantages when they are used in conjunction with screen-printing or similar technologies. The use of an adhesive hydrogel pad dispenses with the need of the standard gel-impregnated sponge, gel- retaining ring, and surrounding disk of adhesive foam that are used in wet-gel electrode designs. It is possible to construct thin, lightweight, highly flexible electrode arrays with accurately defined electrode/gel areas, shapes, and interelectrode distances (72,73).

Hydrogels also tend to cause less skin irritation compared with wet gels. A simplistic explanation of the advantageous/disadvantageous features of hydrogels is that hydrogel serves principally to ensure a good electrical contact between the skin and the electrode and that they do not significantly affect (compared with wet gels) the properties of the stratum corneum.

The impedance of the gel layer can be represented by a simple resi stance in series with the impedances of the skin and the electrode plate–electrolyte interface. The magnitude of the gel resistance will depend on the composition and concentration of the gel and on the dimensions of the gel layer. Hydrogels are generally more resistive than wet gels. Typical resistivities for wet gels are of the order of 5–500 Ω·cm (the higher the salt concentration, the lower the

resistivity) compared with 800–8000 Ω·cm for hydrogels (the higher resistivity hydrogels tend to be used in cardiac pacing electrodes). Wet ECG electrodes, for example, have a gel layer thickness of around 0.3 cm and typical areas of 3 cm^2. The resistance of a wet gel layer is, therefore, generally in the range 0.5–50 Ω. Although hydrogels have higher resistivities, this disadvantage is generally compensated for by the use of larger gel areas, which need not necessarily entail the use of a larger overall electrode area as the adhesive hydrogel may not require the use of a large surrounding disk of adhesive foam. Another way to compensate for hydrogel's inherent disadvantage is to decrease the gel layer thickness, a variable generally ignored in electrode design even though it can have a significant effect on electrical performances. Many commercial hydrogels used in biosignal monitoring electrodes have layer thicknesses of around 1 mm (compared with around 3 mm for pregelled wet electrodes) and, coupled with larger areas of around 7 cm^2, can lead to hydrogel pad resistances in the range 10–100 Ω (5).

It is suggested that further improvements can be made to the performances of hydrogel electrodes (and wet electrodes) by the use of even thinner gel layers. It must be borne in mind that gel-layer resistance is not solely determined by the dimensions and properties of the gel pad. When a large area gel pad is used in conjunction with a small area sensor, the dimensions of the smaller sensor will largely determine the magnitude of the gel-layer resistance, the overlapping section of gel pad carrying relatively little current, which is important in both biosignal monitoring and electrostimulation applications.

Hydrogels, being hydrophilic, are used for wound dressings in order to absorb exudate. They are, therefore, poor at hydrating the skin and will even absorb surface moisture. With hydrogel electrodes, R_{SP} is observed to fluctuate with sweat gland activity and the subject's state of mental arousal, decreasing during increased activity and gradually increasing again as the hydrogel absorbs the excess surface moisture (74,75). In contrast, C_{SP} remains relatively constant after a slight initial increase (32).

Hydrogels are therefore not only more resistive than wet gels, but they hydrate the skin less effectively and give rise to higher skin impedances (i.e., higher values of R_{SP} and lower values of C_{SP}). Typical values of R_{SP} for hydrogels can be as high as 15 MΩ·cm^2 compared with a high of 5 MΩ·cm^2 for wet gels (75). Once again, this disadvantage can be overcome, at least partially, by the use of larger hydrogel pad areas. An additional way of increasing the value of C_{SP} is the use of thinner hydrogel pads (32).

Skin Preparation Techniques. In the clinical environment, the skin site is often degreased using an alcohol wipe prior to electrode application, which probably removes some of the loose, outermost cells of the stratum corneum and the poorly conducting lipid substances from the surface of the skin (55). However, the use of alcohol wipes may initially increase the impedance of the skin by dehydrating the outer layers of the skin (76). Motion artifact also may increase initially following application of alcohol to the skin (64). When wet gel electrodes are applied to alcohol-wiped skin, the gel will eventually penetrate the degreased skin more readily once the electrode has been on the skin for several minutes, leading to a more rapid decrease in skin impedance and possibly to a decrease in motion artifact, which may not be the case, however, in the case of hydrogel electrodes, which do not actively hydrate the skin. The use of an alcohol wipe accompanied by vigorous rubbing should result in low initial impedances due to the additional mild abrasion.

A related method of rapidly decreasing skin impedance is to prerub the skin site with a high concentration electrolyte, thus forcing the gel into the outer layers of the skin, resulting as in a significant decrease in R_{SP} (Fig. 12) and an increase in C_{SP}, especially when accompanied by vigorous rubbing. Arbo-prep cream is supplied for this purpose and it is claimed to reduce skin resistance by up to 90% (from 40 or 50 to 4 or 5 kΩ, according to an advertisement). Some commercial gels such as Hewlett-Packard's Redox paste contain abrasives such as crushed quartz, which, when rubbed into the skin prior to electrode application, greatly reduce skin impedance. Such aggressive gels should only be used in short-term biosignal monitoring applications such as stress testing where high quality traces are required.

The outer layers of the stratum corneum can also be removed by rubbing the skin with abrasive pads especially designed for this purpose, which can give rise to a major decrease in R_{SP} (Fig. 12) and an increase in C_{SP}.

Unomedical, for example, markets a small disposable skin preparation abrasive pad that, when adhered to the finger tip, can be used to dramatically reduce skin impedance. A Skin Rasp, which resembles a strip of Velcro, is marketed by Medicotest for this purpose. The Quinton Quick-Prep Applicator, rotates the abrasive center of the Quick-Prep electrodes, causing a marked decrease in skin impedance. ECG electrodes are often supplied with abrasive pads built into the electrode release backing.

In skin stripping, the stratum corneum is progressively removed by repeatedly applying and removing adhesive tape to and from the skin (55,77). Skin stripping can greatly decrease skin impedance as a consequence of a dramatic decrease in the value of R_{SP} and an increase in C_{SP}. As the outermost layers of the stratum corneum are the most resistive, the most significant decrease in skin impedance is achieved with the first few strippings (77). Therefore, no need exists for the complete removal of the stratum corneum, which would obviously be clinically unacceptable due to the discomfort (pain, bleeding, or irritation) caused to the patient during and following the recording. The more the skin is abraded for a given gel composition, the sooner discomfort develops and the more severe the irritation. The level of irritation also varies with the salt concentration and the additives present in the gel.

As pointed out above, abrading or stripping the skin is often used in stress testing to decrease motion artifact (63) as well as the 50/60 Hz noise induced by any mismatch of the contact impedances. High concentration gels are also often used for such demanding applications, rapidly soaking the skin and, thus, effectively removing the source of the problem.

The use of both skin abrasion/stripping and an aggressive gel will, however, maximize the potential for severe

skin irritation problems. These approaches should not be used together. Even with the use of mild gels, it is probably unwise to abrade the skin for long-term monitoring applications. The increased length of exposure of the abraded skin to the gel will be conducive to skin irritation. Somewhat surprisingly, long-term monitoring electrodes are sometimes commercially supplied with integral abrasive pads, which is not only risky but probably unnecessary as the use of a suitable mild gel would eventually decrease the skin impedance without the need for skin abrasion.

ELECTRODE DESIGN

External Biosignal Monitoring Electrodes

Historical Background. In 1887, Augustus Waller, using Etienne Jules Marey's modification of the capillary electrometer, obtained surface ECGs (as opposed to recording directly from the exposed heart of an animal) of one of his patients, 'Jimmy'. The patient turned out to be his pet dog. Waller used two buckets of saline to measure the canine ECG, one for the front paws and one for the hind paws (78–80).

Waller eventually succeeded in recording the first human ECG in 1887 using the capillary electrometer (Fig. 14) (81). However, he initially concluded "I do not imagine that electrocardiography is likely to find any very extensive use in the hospital. It can at most be of rare and occasional use to afford a record of some rare anomaly of cardiac action" (82).

It was, therefore, left to a more visionary and tenacious Dutchman, Willem Einthoven, to establish the clinical relevance of this strange new trace and to develop and commercialize a clinically acceptable system based on the string galvanometer. Einthoven's achievement was truly awesome. However, it must be pointed out that he did build (very significantly, it is conceded) on the work of earlier pioneers. The electrode system used, for example, was

Figure 14. Human subject connected to capillary electrometer via large area bucket electrodes (81).

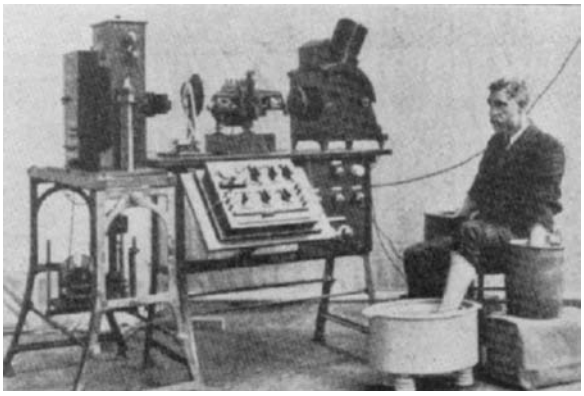


Figure 15. Early commercial ECG machine and electrodes (84).

Waller's bucket electrode, whereas the moving photographic plate recording technique was originated by Marey (83).

The input impedance of Einthoven's galvanometer was such that very low contact impedances were necessary, hence, the very large bucket electrodes (Fig. 15) (84). Obviously, the range of applications was somewhat limited.

Realistically, only one's limbs could be conveniently placed into the buckets. Hence, the use of limb leads exists in electrocardiograpghy, even to the present day. It is, therefore, important to note several points. Present state-of-the-art is often based on historical quirks rather than on a profound scientific basis. The monitoring device and amplifier determined the electrode size, design, and location of the electrodes, which in turn determined the clinical application and the presentation of the physiological data.

Einthoven's device in its early form could not be used for the monitoring of bed-ridden patients or for ambulatory monitoring. These applications had to wait for improvements to be made to the amplifiers, which then enabled the use of smaller electrodes that could be more conveniently attached in other anatomical locations. However, the early monitoring locations and the form of the signals observed became accepted as standard and there is often considerable resistance to novel monitoring scenarios (e.g., smart clothing), which require or are based on different lead systems and present physiological data in a different format to that familiar to the clinician.

In the 1920s vacuum tubes were used to amplify the electrocardiogram instead of the mechanical amplification of the string galvanometer, which lead to smaller, more rugged systems that were transportable (Fig. 16) (84). The input impedances of the new ECG monitors were larger, and the large metal buckets could be replaced by smaller metal-plate electrodes (still large compared with present-day electrodes) (83). These advances enabled bedside monitoring, and, by the 1930s, some ECG devices could be carried to the patient's home. Not unsurprisingly, the new plate electrodes were attached to the limbs, both for historical and practical reasons. The metals used were chosen for their availability and ease of machining (Fig. 17). They included German silver, nickel-silver,

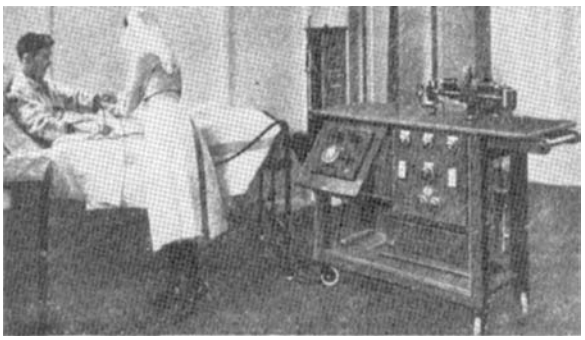


Figure 16. Mobile ECG used to monitor bed-ridden patient in hospital ward circ. 1920 (84).

and nickel-plated steel. The foil plates were used in conjunction with moistened pads of paper toweling, lint, cotton gauze, or sponge and were generally held in place with rubber straps. Around 1935, conductive gels were developed to replace the soaked pads. A wide range of gel ingredients were assessed, and it was noticed that the presence of an abrasive in the gel greatly reduced the skin impedance (5).

They also noted that slightly abrading the skin before applying the electrolyte helped achieve very low skin impedances.

A dry version of the plate electrode was reported by Lewes (85). The multipoint stainless-plate electrode resembled a large nutmeg grater that penetrated the skin when firmly strapped onto the limb or applied to the skin with a slight rotary movement, thus resulting in a very significant reduction in skin impedance.

Modern versions of the limb plate electrode still exist. Some have a convenient spring clip mechanism, which dispenses with the need for the rubber strap.

In the 1930s, clinicians, some using electrodes held on the chest by the patient himself or by another member of clinical staff, experimented with precordial leads and established their clinical value (86). In 1938, the American Heart Association and the Cardiac Society of Great Britain defined the standard positions and wiring of chest leads V1–V6 (87).

Research then focused on the development of electrodes that could be conveniently attached to the chest to enable convenient routine clinical measurements. Several designs

Figure 17. Metal plate limb electrode.

involved a rubber bulb, which was used to create suction sufficient to hold the metal electrode on the chest. One of the first suction electrodes was developed by Rudolph Burger in 1932 for the precordial leads (88). The suction electrode shown in Fig. 18a (85,89) is one developed by Ungerleider (89). Another more recent system incorporated the multipoint electrode of Lewes (85) into the suction head (Fig. 18b). The most popular suction electrode design, widely used around the world and still in use today, was developed by Welch (90) and often called the Welch or Suction cup/bulb electrode (Fig. 19) (3). It consists of a hollow, metallic, cylindrical electrode that makes contact with the skin at its base. A rubber suction bulb fits over the other end of the cylinder. The suction bulb was squeezed while the electrode was held against the skin. Upon releasing the bulb, the electrode is held in place. The suction electrode can be used anywhere on the chest and can even be used on hairy subjects. A single electrode can, if necessary, be used to take a measurement at a given location and then moved to another site.

Although the Welch cup electrode became widely used as a precordial electrode, it could only be realistically used for resting (supine) diagnostic ECG recording. The weight and bulk of the electrode generally rules out its use on upright, ambulatory, or clothed subjects. Since then, more suitable, lightweight, low profile suction electrodes have been developed that are pneumatically connected to remote vacuum pumps (37). Some arrays of suction electrodes are commercially available, for example, for use with exercise bicycles for cardiac stress testing (72).

A method had to be invented to attach small disks of suitable metal and their conductive gel coating to a patient's chest (in the case of ECG) or to other body parts in the case of other biosignal applications, such as EEG and EMG. Simply taping a metal disk to the skin site with a sandwiched gel layer was a method often used (91).

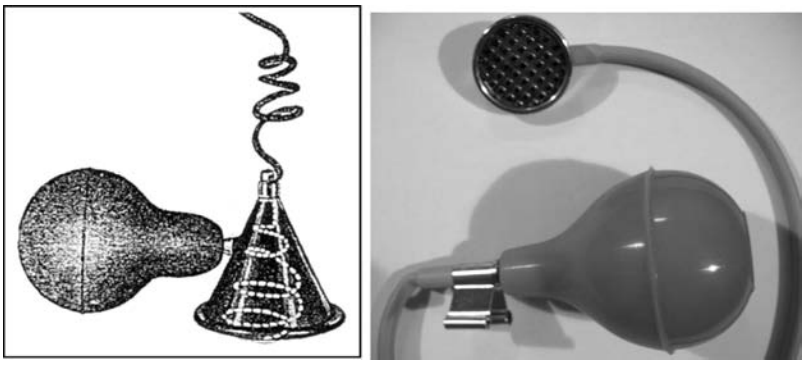


Figure 18. Early designs of suction precordial electrode (a) Ungerleider (89). (b) Lewes (85).

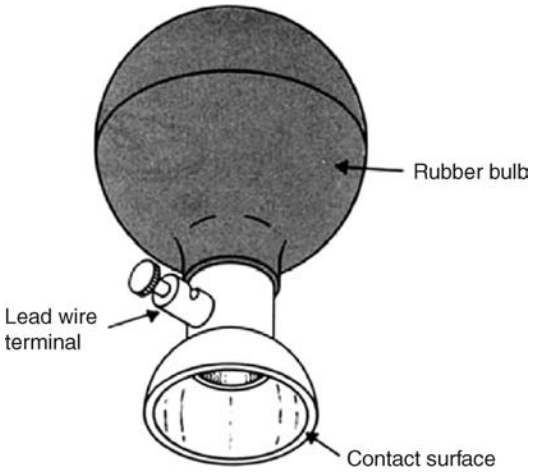


Figure 19. A metallic suction electrode is often used as a precordial electrode on clinical electrocardiographs (3).

Conically formed metal disk electrodes were and often still are used for EEG recordings (Fig. 20). The base of the metal cone is attached to the patient's scalp using elastic bandages, wire mesh, or more recently, using a strong adhesive such as colloidon. Aperture exists in the apex of the cone to enable the introduction into the recessed electrode of electrolyte gel or to enable the abrasion of the underlying skin by means of a blunt hypodermic needle. The cone electrodes were often made of gold as it has high conductivity and inertness, desirable in reusable electrodes. More recently, Ag/AgCl has been used.

Early plate electrode designs were presumably very messy and gave rise to considerable artifact problems. The observed artifacts were attributed to disturbance of the double-layer region at the electrode/skin (or, more precisely, electrode/electrolyte) interface [termed the electrokinetic effect by Khan and Greatbatch (94)]. When the electrode moves with respect to the electrolyte, the distribution of the double layer of charge on electrode interface was thought to change and cause transient fluctuations in the half cell potential or give rise to a streaming potential.

Recessed or floating electrodes were introduced in an effort to protect the electrode–gel interface from such mechanical disturbance and resultant movement artifact.

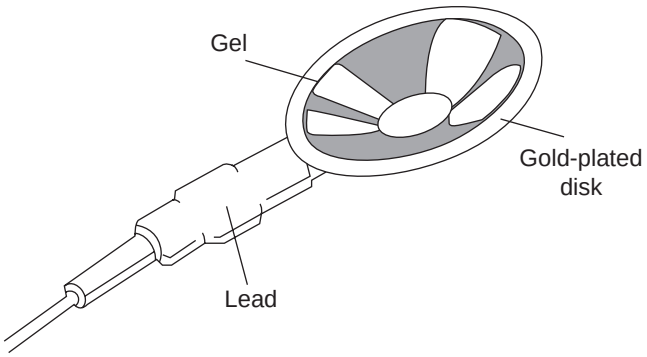


Figure 20. Conically formed metal disk EEG electrodes.

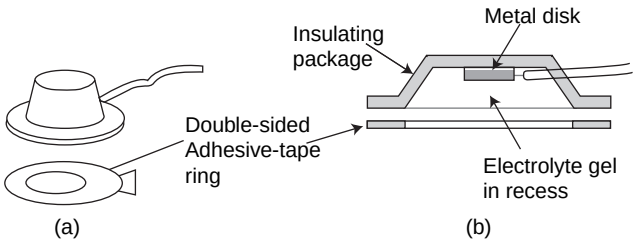


Figure 21. Examples of a floating or recessed biosignal electrode. (a) Recessed electrode with top-hat structure. (b) Cross-sectional view of the electrode in (a)(3).

A metal disk was recessed in a plastic housing that was filled with electrolyte gel prior to application to the patient. The top hat-shaped container was adhered to the skin by means of an annulus of double-sided adhesive tape, (Fig. 21) (3). Later a gel-impregnated sponge was used to ensure good electrical contact between the electrode disk and the skin surface. The electrode disk was, therefore, not in direct contact with the skin, which was found to reduce motion artifact. At first, various metal plates were used as the electrode conductor, then a sintered Ag/AgCl disk with preattached wire ensured better performances for more demanding applications.

Modern Disposable Electrodes. The top hat housing was eventually replaced with a smaller retaining ring or plastic cup and the electrode was held in place by means of a surrounding disk of adhesive foam. The plastic cup holds the gel-impregnated sponge in place and stops the gel from spreading beyond the set boundary, either during storage or use on the patient. Low cost Ag/AgCl-plated plastic eyelets (part of a snap fastener) are used in these disposable electrodes and the leads are connected to the electrodes via the electrodes' snap fastener studs. The rigid retaining ring was, however, uncomfortable as it did not allow the electrode to conform optimally to body contours. It was eventually removed in many modern disposable electrodes and the recess is now often formed by a hole in the adhesive foam layer. The backing label serves to hold the snap and eyelet in place as well as to present the company's logo (Fig. 22). The resultant electrode structure is much more flexible and more comfortable to wear.

The use of a snap fastener-style connection in disposable electrodes has one significant drawback for certain applications. The male stud protruding from the back of the electrode and the female connector required on the connecting lead results in a relatively heavy, large-profile electrode/connector interface, which is less than optimal for applications such as neonatal and pediatric monitoring. The use of such electrodes in long-term monitoring of bedridden patents could lead to considerable discomfort and the heavy connection could also give rise to significant motion artifact problems. The integrated lead design seeks to overcome these disadvantages. A thin, highly flexible lead wire is bonded directly to the back of a specially designed Ag/AgCl-coated eyelet, which results in a very low profile, lightweight electrode-connector system much used in neonatal and pediatric monitoring and attractive for long-term monitoring applications (Fig. 23).

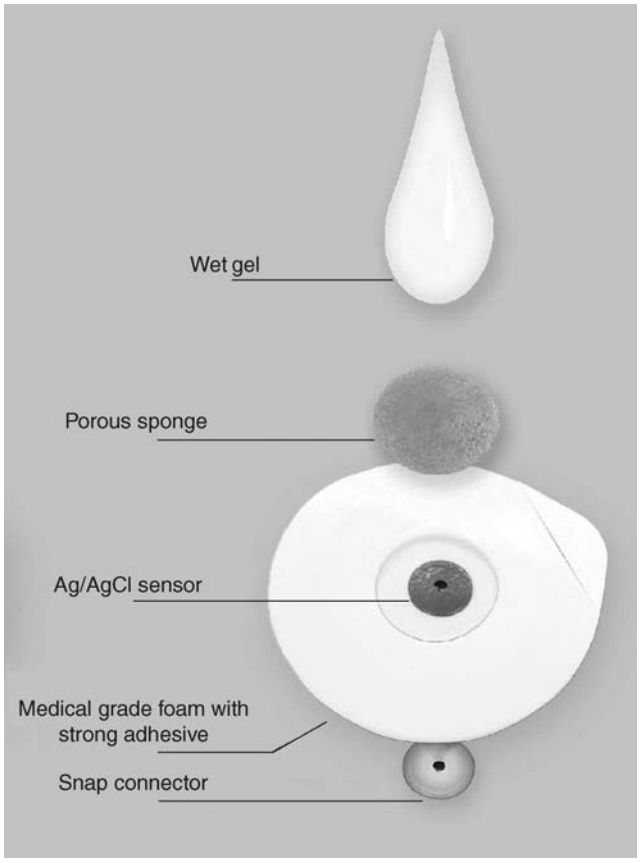


Figure 22. Modern wet gel disposable electrode. (Courtesy of Unomedical A/S.)

In an effort to decrease motion artifact, many electrode designs feature an offset center. The connector, often in the form of a snap fastener, is separated from the gelled sensor by a strip of metal or similar conductive layer. The connector is thus 1 or 2 cm away from the metal–gel–skin interface, and it is possible to connect the lead to the electrode or to pull on the connector without pulling directly on the gelled, skin site, thus causing artifact problems. This design appears well suited for stress testing applications although arguably less so for long-term monitoring of bed-ridden patients due to the bulky connector. The invention was patented by Manley (93) and the concept

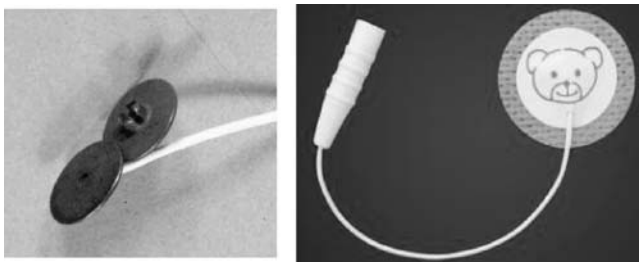


Figure 23. (a) Specially designed Ag/AgCl-coated eyelet with bonded lead. (b) Low profile, lightweight pediatric electrode with bonded lead. (Courtesy of Unomedical A/S.)

Figure 24. An example of an off-set connector electrode. (Courtesy of Unomedical A/S.)

has been commercially exploited very successfully by Ambu A/S.

More recently, many other manufactures supply electrodes with offset connectors (Fig. 24). Leadlock have developed an electrode that incorporates a slit in the foam backing. Once the electrode has been applied to the patient and the lead connected to it, part of the foam backing is used tape down the lead, locking it in place and minimizing direct pull on the connector and underlying electrode/skin interface.

A wide range of backing materials now exists, and some are better suited for specific monitoring applications, types of patients, skin types, and so on. Given the great variety of materials, adhesives, and designs used, the following comments are generalizations.

Open-cell foam layers, made from a plastic such as polyethylene, are much used and have thicknesses typically in the range 1–2 mm. They have a 10–100 μm coating of a pressure-sensitive adhesive, generally a polymeric hydrophobic substance (68,94). Adhesive foams generally give rise to firm adhesion, are resistant to liquids, and tend to cushion lead pull, thus giving rise to less artifact. They are, therefore, generally well-suited to cardiac stress testing and similar applications. However, as they are occlusive and generally have a relatively aggressive adhesive, they can give rise to more skin irritation and they must be used with caution for neonatal and long-term monitoring.

Porous, breathable layers, such as nonwoven clothes or tapes, have the advantages of being soft, stretchable, and conformable with the skin. (*Note:* The term micropore, although sometimes used loosely to describe any breathable backing material, is strictly a 3M product.) Porous layers tend to cause less mechanically induced trauma to the skin, which can occur with more rigid materials and is due to shearing of underlying skin layers. As they are highly air permeable and use milder adhesives, porous tapes cause less skin irritation and are well suited for long-term monitoring. Larger backing areas tend to be required. The gelled center can, however, pull away from the skin as a result of the stretchable backing. Ambu A/S

use a central ring of adhesive around the gelled eyelet to minimize this problem.

As we have already seen, wet (as opposed to solid gels) vary in composition and concentration depending on the application. Aggressive gels with higher concentrations of electrolyte or including abrasive particles are used for short-term, demanding monitoring applications such as cardiac stress testing. Mild gels are used in pediatric and neonatal applications due to the increased vulnerability of the patient's skin. It should be noted that no matter how hypoallergenic a gel or an adhesive is claimed to be, some patients will experience some form of skin reaction to one of the components.

Solid Conductive Adhensive Electrodes. The growing monitoring market has led to the development of even lower-cost disposable electrodes. Solid conductive adhesive or hydrogel electrodes were first introduced by LecTec Corp around 1980 (95). Hydrogels are composed of a hydrocolloid, alcohol, a conductive salt, water, and a preservative. The hydrocolloid use can be either natural (e.g., Karaya gum) or synthetic (e.g., polyvinyl pyrrolidone) (68). Early hydrogel electrodes were based on the natural hydrocolloid, Karaya, which comes from the bark of a tree. The rather unaesthetic appearance of these early gels and the variations in their electrical and mechanical properties limited their widespread acceptance. The use of synthetic hydocolloids, with their more attractive appearance and performances, has led to the recent revolution in electrode design.

Solid adhesive gels reduce the number of electrode parts required, dispensing with the need of a gel-impregnated sponge or a surrounding disk of adhesive tape, which gives rise to small-area, low profile electrodes suitable for neonatal monitoring, especially when coupled with integrated leads as discussed above (Fig. 23b).

Tab solid adhesive electrodes are now widely used for many biosignal monitoring (and stimulation) applications. Thin, highly flexible metallic/conductive foils or printed conductive ink layers are laminated with solid, adhesive hydrogels. A section of the foil or printed layer is left uncovered. Once the electrodes are cut out, the exposed conductive tab acts as a means of connection, the leads being connected via alligator clips (Fig. 25). Electrode design is therefore very simple and manufacturing costs are low. These flexible, low profile electrodes are best used for short-term, resting diagnostic monitoring. Tab electrodes are not suitable for ambulatory or long-term monitoring as the tab connection will cause the electrode to peal off quite easily when pulled from any angle other than directly downward. Also, hydrogels are hydrophilic and tend to absorb moisture, lose their adhesive properties over time, and fall off the patient if an additional adhesive backing is not used. Hydrogels, being solid, do not leave a messy residue on the skin requiring cleaning. Tab electrodes are also repositionable and are reuseable (on the same patient!) in certain home monitoring applications.

When used with an adhesive backing layer, the hydrophilic hydrogels tend to be relatively nondrying (a significant problem with pregelled wet electrodes) and their electrical properties may even improve as they absorb moisture. As they do not actively hydrate or otherwise affect the skin, they tend to be relatively nonirritating compared with wet gels.

Figure 25. Hydrogel-based tab electrode with connector. (Courtesy of Unomedical A/S.)

Some disadvantages exist, however, associated with hydogels. Hydrogels are more resistive than wet gels and, hence, the gel pad resistance will be higher, which can be compensated for by using larger hydrogel pad areas and thinner layer thicknesses as compared with those used with wet gels. Although the area of the solid adhesive gel in a tab electrode, for example, is considerably larger for this reason than that in a standard disposable wet gelled electrode, the absence of a surrounding adhesive layer results in the tab electrode having a smaller overall area.

Hydrogels, being hydrophilic, are poor at hydrating the skin and may even absorb surface moisture. They, therefore, give rise to larger skin impedances. This disadvantage can also be overcome, at least partially, by the use of larger hydrogel pad areas. Hydrogels are also more expensive than wet gels but generally lead to less expensive electrodes due to the simpler designs involved.

Hydrogels are more sensitive to motion artifact as they do not actively hydrate the skin. They are, therefore, not well-suited for stress testing.

The use of such solid gels entails numerous advantages when they are used in conjunction with screen printing technology (73,96), especially for body surface mapping and similar applications. It is possible to construct thin, lightweight, highly flexible electrode arrays with accurately defined electrode/gel areas, shapes, and interelectrode distances for a wide range of novel stimulation and biosignal recording applications. As the solid gel will not spread between electrodes, it is possible to position electrodes very close together without electrical shorting (Fig. 26).

Wearable Electrodes for Personalized Health. The recent and continuous trend toward home-based and ambulatory monitoring for personalized healthcare, although exciting and potentially leading to a revolution in healthcare provision, necessitates even more demanding performance

Figure 26. Cardiac mapping electrode harness.

criteria for the monitoring sensors (97,98). Many groups around the world are seeking to incorporate electrodes into clothing in order to monitor military personnel, firefighters and eventually the average citizen who wishes to monitor his or her health. Systems already exist on the market (e.g., Life Shirt) that resemble waistcoats into which one plugs-in standard ECG electrodes and other sensors. These sensors are removed and replaced periodically by the subject and, hence, require the knowledgable involvement of the motivated wearer, presently military personnel, athletes, rescue workers, and so on.

For the more widespread use of wearable monitoring systems, especially by the average citizen, the system must be very easy and comfortable to use and require no preparation—literally as simple as putting on their shirt. Electrodes must, therefore, (*1*) require no prepping, (*2*) be located in the correct location once the smart garment is put on, (*3*) make good electrical contact with the skin, (*4*) not give rise to motion artifact problems, (*5*) not cause discomfort or skin irritation problems, and (*6*) be reusable and machine-washable. Although much work has been carried out in this novel area, it is not surprising given the above list of required performance criteria that the electrodes/sensors tend to form the bottleneck in the success of the overall monitoring systems. One must, therefore, not simply choose an electrode with as conductive a metal element as possible. Unfortunately, it would often appear that the associated electronic systems are first developed and the electrode design is left to the end, almost as an afterthought. The author would, therefore, suggest that researchers start with the desired biosignal and establish the optimal body site(s) and electrode design for the given application before developing the rest of the monitoring system. This process may involve the use of novel lead or montage electrode positions in order to conveniently pick up artifact-free signals. Although this method will necessitate clinicians interpreting nontraditional waveforms, it will at least enable feasible monitoring and, as it involves novel body sites and electrode designs, it may well be patentable. After all, if it is not patented and commercialized, it will not benefit the patient.

One of the most promising smart garments is that developed under a European Fifth Framework programme called WEALTHY (Wearable Health Care System) (Fig. 27). WEALTHY is a wearable, fully integrated system, able to monitor a range of physiological parameters including electrocardiogram, respiration, posture, temperature, and a movement index. Fabric electrodes are made using conductive fibers woven into the stretchable yarn of the body-contour hugging garment and connections are integrated into the fabric structure (Fig. 27b). Various membranes are being assessed to ensure optimal electrode-skin contact and minimize skin irritation. The garment is comfortable and can be worn during everyday activities. It is washable and easy to put on.

External Electrostimulation Electrodes

Historical Background. The evolution of external stimulation electrode design shares some of the key landmarks as the development of biosignal monitoring electrode and, hence, this section will be somewhat shorter.

From the mid-1700s, when electrostatic generators were used to deliver arguably therapeutic impulses to various parts of the body, handheld (by the practitioner) electrodes had to be designed capable of delivering the impulses to the patient without shocking the practitioner who was holding them against the body part in question. The electrodes used tended to be simply long metal rods insulated with wooden handles (Fig. 28) (99). Although the electrodes were initially terminated in a simple metallic sphere, more exotic terminations were soon invented as these were observed to lead to different therapeutic effects on the body by means of the variations in the streams of the electric fluid. A modern parallel would be the use of different pencil electrodes (ball, loop, and needle) in electrosurgery for different effect.

Figure 27. (a) The WEALTHY physiological monitoring vest with integrated sensors. (b) An early version of the fabric electrodes.

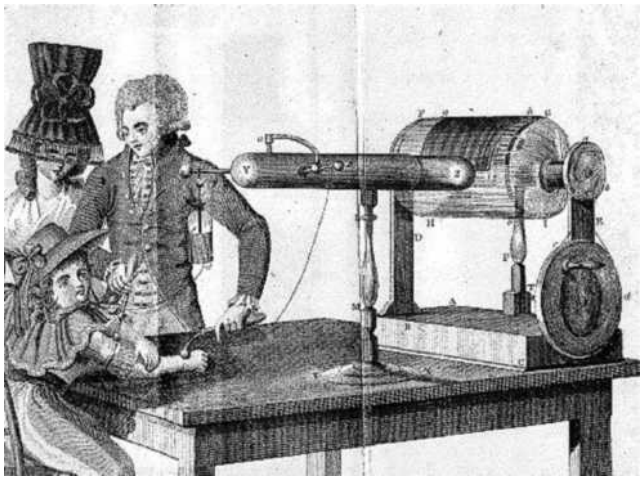


Figure 28. Early electrostatic generator and handheld electrodes (99).

Around 1800 came the discovery of Galvanism (dc current) and Voltaic Piles (early batteries). Numerous examples exist of practitioners using their handheld probe for localized effect, and the second contact to the patient was made by means of a container of water into which the patient put a hand or foot (Fig. 29) (100).

Following the discoveries of self-and mutual induction (~1830), Guillaume Duchenne made great contributions to the clinical application of the new Faradic current. At that time, much interest existed in the localization of what became known as motor points. It was common to combine the prevailing interest in acupuncture and use needles to stimulate muscles and nerves under the skin, termed electropuncture (83). Duchenne was not happy with this approach and developed his own electrodes for localized electrization. His electrodes were in various shapes (disks spheroids, and cones) covered with leather

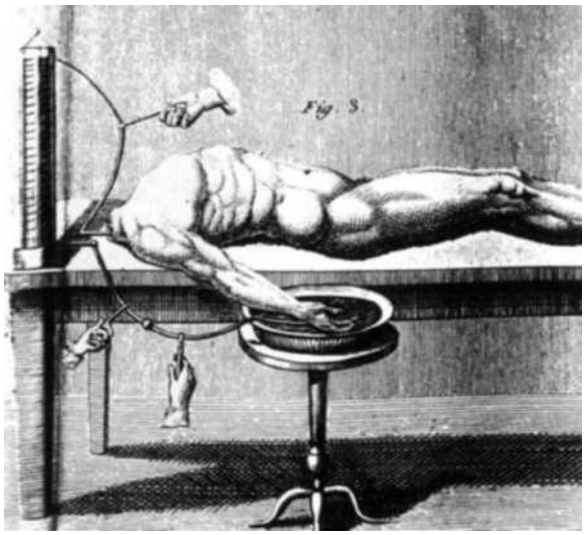


Figure 29. Container electrode. Used on the dead and the living (100).

Figure 30. Duchenne's moistened conical electrodes for localized electrization (101).

moistened with salt water prior to application (101). (Fig. 30).

During the 1900s, as with biosignal monitoring electrodes, electrostimulation electrodes involved the use of simple metal buckets or receptacles, filled with water or another electrolyte, into which the subject introduced their foot or hand, especially in early iontophoretic applications. Obviously, the range of applications was somewhat limited. Metal probes were still manually pressed against skin for short-term applications. The electrodes were either gelled before application or had moistened chamois coverings similar to those used by Duchenne. In the 1950s, early external pacing and defibrillator electrodes, termed paddles because of their shape, consisted of bare metal disks made of noncorrosive material and were simply pressed against the patient's chest (102) see Fig. 31.

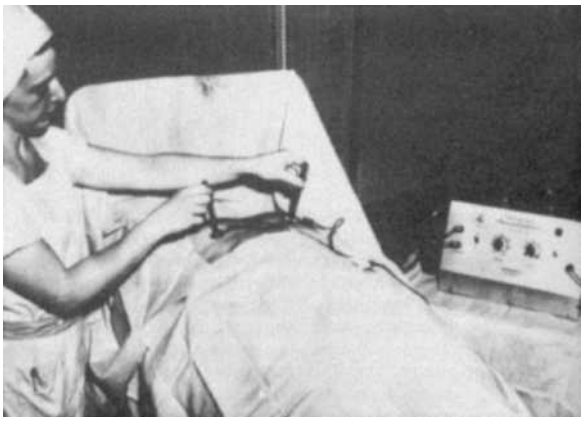


Figure 31. Early pacing equipment and handheld electrodes (102).

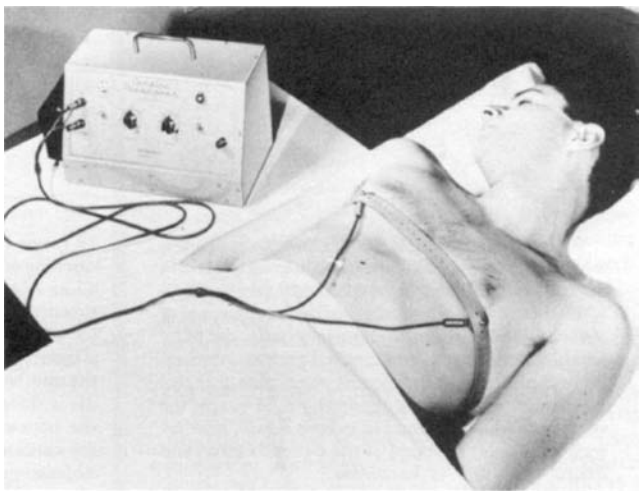


Figure 32. Early pacing equipment and metal-plate electrodes (103).

Rigid metal plate electrodes were eventually held in place with rubber straps on the limbs and even the thorax (Fig. 32) (103). Some of these electrodes were and still are made with rigid stainless-steel plates (104). The foil plates were generally used in conjunction with moistened pads of paper toweling, lint, cotton gauze, or sponge. The pads were moistened by the therapist prior to electrode application with water or electrolyte. Such electrodes could be easily reused by simply washing and regelling the electrode. Being rigid, however, these plate electrodes did not always make optimal contact with the body surface and gave rise to current density hot spots. External cardiac pacing at this time, for example, was very painful (83).

Malleable metal foil electrodes were the next evolutionary step in electrode design. Malleable electrodes have been made using a range of metals including tinplate lead and aluminium foils (105). Such electrodes had the advantage of being able to conform, to some extent, with body contours, thus ensuring a better, more comfortable contact between the electrode and the patient than was the case with rigid plates. Wrinkles in malleable metal foil could, however, encourage preferential current flow through small areas of the gel and into the patient.

More convenient, disposable pregelled foil electrodes were then developed for a range of external electrostimulation applications. The metal foil was laminated onto an adhesive foam backing. A gel-impregnated sponge layer was located on top of the metal layer and the complete electrode is attached to the patient by means of the surrounding layer of adhesive backing foam.

Unfortunately, the wet gel in these disposable pregelled electrodes tended to pool to one side, depending on how they were stored, giving rise once again to current density hotspots. More recently, the gel-impregnated sponge layer has been replaced by a conductive adhesive gel layer, as it does not have the potential for pooling to one area during storage and it does not squeeze out under pressure (68) (Fig. 33).

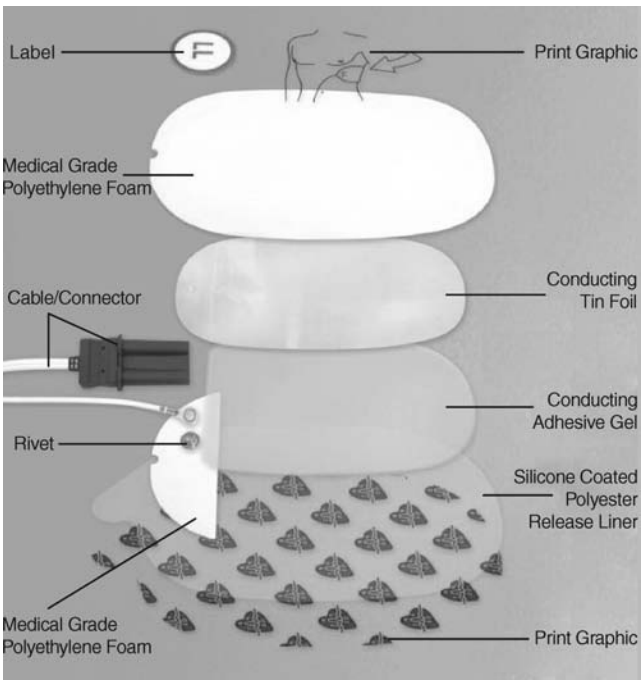


Figure 33. Construction of a modern external pacing or defibrillation electrode (courtesy Unomedical A/S).

Current Density Considerations. The distribution of current density under an electrode is an important parameter when designing and using electrostimulation electrodes. In the simplest case, current density (the amount of current per unit of conduction area) is inversely proportional to the electrode/skin contact area. For a given current, the current density under a small-area electrode will be higher and more localized than that under a large-area electrode. Generally, an optimal electrode area exists for a given therapeutic application, based on a range of criteria including the anatomical position and size of the nerve/muscle/organ and the relative positions and sizes of the electrodes. Small electrodes are, therefore, well existd to target precisely known points such as motor points. If a small electrode is used in conjunction with a large-area electrode, the effect is more pronounced under the smaller of the two. In such monopolar stimulation, the small electrode is often used as the active electrode to target the therapeutic effect. The larger electrode is simply used to complete the electrical circuit and is termed the indifferent or dispersive electrode. The use of two equally sized electrodes is termed biopolar stimulation. In TENS, for example, biopolar stimulation is often used to stimulate large muscle groups sandwiched between the (large) electrodes. Too large an area of electrode, however, may cause the current to spread to neighboring tissues.

High current densities can cause tissue injury due to, among other things, heating effects. The passage of electricity through any conductor will cause the dissipation of heat within that conductor. The amount of heat generated in a tissue depends on spatial and temporal patterns of current density and tissue resistivity (49).

The total energy dissipated at an electrode–skin interface is given by the formula:

$$E = I^2Rt \tag{21}$$

where

E is energy dissipated (J)
I is root-mean-squared (rms) electrode current (A)
t is the duration of current flow (s)
R is the real part of the impedance at the electrode site (Ω).

The change in temperature at the skin, ΔT, site is proportional to the energy dissipated and, hence, ΔT is proportional to I^2Rt. When skin or muscle tissue is heated to about 45 °C for prolonged periods, thermal damage can result. For short durations (i.e., <5 s), a temperature rise approaching 70 °C would be needed to cause heat damage.

As the electrode–skin resistance, R, is not generally known for a given site, it is often found convenient to use a heating factor (HF), where

$$\mathrm{HF} = I^2t \quad (\mathrm{A}^2 \cdot \mathrm{s}) \tag{22}$$

Assuming uniform current density distribution under an electrode, it is possible to calculate the minimum area of electrode necessary to achieve therapeutic effect and avoid tissue trauma (42). In theory, the applied currents flowing through standard dispersive electrodes used for electrosurgery, for example, will generally not give rise to sufficiently high overall current densities to cause thermal damage. However, analysis shows that current density distribution is not uniform under a stimulation electrode and that localized hotspots can occur and cause considerable pain and trauma to the patient when applying apparently safe' therapeutic impulses (49). At best, in cases such as TENS, the applied current may have to be limited to less than therapeutic values due to the patients discomfort (68).

Many potential sources exist of accidentally high current densities. Wrinkles or breaks in the metal electrodes, gel squeezing out from under the electrode or drying out, electrodes partially peeling off the skin, poor electrode application, and so on can encourage preferential current flow through small areas of the gel and into the patient. However, current density hotspots can also occur due to poor electrode design, and a considerable amount of research has been and is being spent investigating this important problem.

In this presentation, stimulation electrodes have been divided into conductive electrodes and resistive electrodes in order to facilitate the review of the various design features.

With highly conductive metal electrodes, such as those used for external cardiac pacing, defibrillation, or electrosurgery, current density hotspots are observed to occur under the perimeter of the electrode, often evidenced in the past by annular-shaped burns to the patient (49,106).

Current density hotspot problems are now often studied using thermal imaging cameras. Thermograms of the patient's skin (or a substitute such as pig skin) are taken immediately following the application of a given series of pulses and the removal of the electrode under test (Fig. 34). Increases in skin temperature reflect the magnitude of the current density at a particular point (107).

Wiley and Webster (108) showed that current flow through a circular electrode placed on a semi-infinite medium could be solved analytically. They found that for an electrode of radius, a, and total current, I_0, into the electrode, the current density into the body as a function of radial distance from the center, r, was given by:

$$J(r,0) = \frac{J_0}{2[1-(r/a)^2]^{1/2}} \quad (\mathrm{A} \cdot \mathrm{cm}^{-2}) \tag{23}$$

where $J_0 = I_0/\pi a^2$, (i.e., a hypothetical uniform current density).

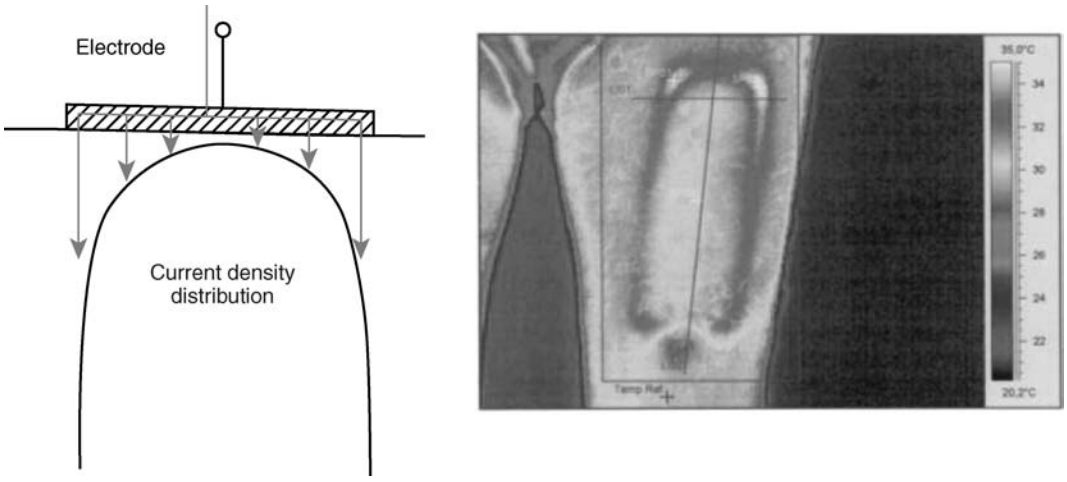


Figure 34. (a) Schematic representation of the current density distribution under a conductive electrode plate. (b) Thermal image of the skin under an electrosurgical electrode following testing.

As a result of the approximations made in deriving this simple equation, the value of current density at the edge (when $r = a$) would theoretically approach infinity. More realistically, the current density at the perimeter can be around three times higher than that at the center of the electrode (109) (Fig. 34). The above equation shows that the middle portion of the electrode is relatively ineffective in carrying the current as half the total current flows through an outermost annulus 0.14 a wide or one-seventh of the radius.

Efforts in this area concentrate on encouraging more of the current to flow through the central portion of the electrode.

A main concern in the design of the conductive stimulation electrodes used for external cardiac pacing, defibrillation, or electrosurgery is the decrease in the high current densities observed at the edges.

In TENS, relatively resistive conductive rubber is often used and the opposite problem develops. When current is introduced into the conductive rubber (via a small metallic connector), it tends to flow into the skin immediately under the connector rather than laterally through the resistive electrode. Efforts in this area concentrate on encouraging the current to flow laterally through more of the electrode surface.

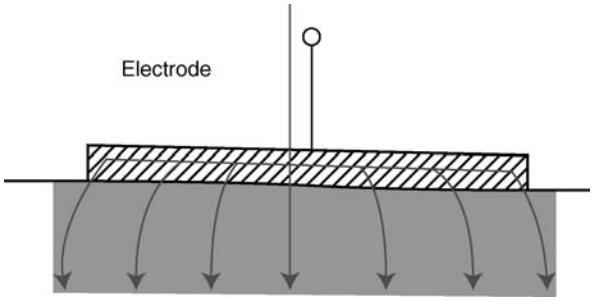


Figure 35. Schematic representation of the current density distribution under an electrode plate coupled with a larger gel pad. Current is allowed to spread out beyond the boundaries of the metal plate, thus minimizing the edge effect.

Modern Electrode Designs

Conductive Electrodes. Electrosurgery, external cardiac pacing, and defibrillation share a common problem: Electrodes tend to deliver or sink a substantial portion of the outgoing or incoming current through their peripheral area as opposed to providing a uniform current density along their surface. This problem is referred to in the literature as the fringe, edge, or perimeter effect.

Many suggestions have been made to reduce this edge effect observed with metal electrodes, including:

1. Increasing overall area of the electrode. Obviously, an increase in electrode area will lead to a decrease in current density (110,111). However, it is generally not practical to use very large electrodes as the applied electrical field must be sufficiently focused to stimulate the targeted tissues and them alone. Also, a strong commercial interest exists in decreasing the size of the electrodes to save money and to facilitate packaging and storage of the electrodes.
2. Avoiding sharp edges in the metal plate (110,111). It has long been observed that square or rectangular electrodes with angular edges concentrate the electrical field at their corners, giving rise to current density hotspots in these locations. Using round electrodes or rectangular ones with rounded edges have been found advantageous in this regard.
3. Making the gel pad slightly larger than the electrode to enable the electric field lines to spread out before reaching the skin (111) (Fig. 35). Using a gel pad much larger than the size of the metal plate has less effect than would be expected as the perimeter of such a large gel pad will carry little current and the additional gel is electrically redundant, which arguably applies to some of the snap connection electrodes used in TENS that do not have an additional current dispersing element.
4. Increasing the overall resistance or thickness of the gel layer in order to give the current more time to spread out evenly through the gel (110,111). It is well-known that, in applications such as external cardiac pacing, the use of relatively resistive gels decreases the pain and burning to the patient's chest. Krasteva and Papazov (110) suggest that the use of a layer of intermediate resistivity, comparable with that of the underlying tissues, optimally improves the distribution. However, in other applications such as external defibrillation, a high resistance gel pad would lead to energy wastage and a decrease in the desired therapeutic effect. Taken to its logical conclusion, this approach results in the coating of the electrode metal plate surface with a dielectric film. Such capacitive electrodes have been shown to give rise to nearly uniform current densities (107).
5. Increasing the resistance or thickness of the gel at the edges. Kim et al. (112) proposed covering the electrode metal with resistive gel of increasing resistivity as one moved out from the center toward the periphery, according to a specific relation with respect to the electrode radius. Although an intriguing concept, the commercial manufacture of such an electrode system is not yet feasible.
6. Making the electrode conductive plate progressively more resistive toward the peripheral edge of the electrode. Wiley and Webster (108) suggested subdividing the electrode plate into concentric segments and connecting external resistors to the individual segments. The connected resistors had progressively higher resistances toward the periphery in order to equalize the currents in the separate segments. A simpler system that has been successfully commercialized was patented by Netherly and Carim (113). A resistive layer is deposited on the outer edge of the electrode conductive plate, thus forcing more current to flow through the central portion of the electrode (Fig. 36). Krasteva and Papazov (110) demonstrated theoretically that a high resistivity perimeter ring

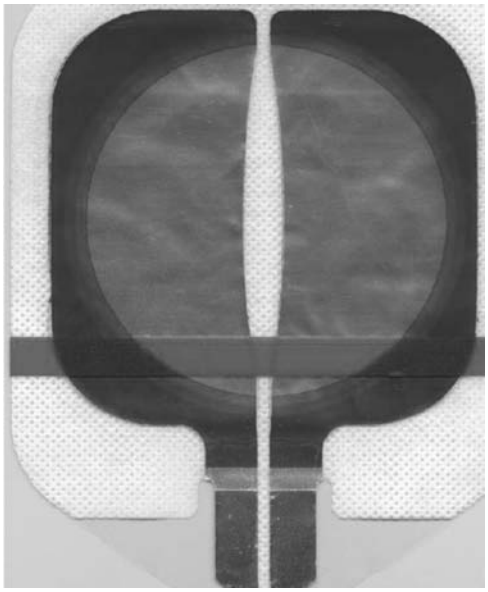


Figure 36. 3M's electrosurgical dispersive electrode. Note the green lossy dielectric material deposited around the peripheral edge of the electrode.

decreased the maximum periphery current by 12% without increasing the total interface resistance, hence the resistance to the defibrillation current.

Another related approach to this problem starts with a resistive graphite-based conductive layer and progressively builds up multilayered (at least two) coatings of more conductive silver/silver chloride toward the center of the electrode (114). The perimeter resistance is approximately 200 times that of the center, this is the technique exploited in Medtronic EDGE system electrodes for defibrillation, noninvasive pacing, and ECG monitoring (Fig. 37). It is claimed that the design distributes the current density evenly over the entire surface area of the electrode, rather than concentrating it at the edges.

7. Scalloping or otherwise shaping the edges of the metallic plate so that the length of the perimeter is increased and, hence, the peripheral current density is decreased. Over the years, various designs have incorporated this concept. For example, it has been shown that using a figure-eight design rather than a rectangular metal plate reduced the maximum temperature (reflective of current density) by 30–50% (107). An alternative design is shown in Fig. 38. Caution is advised with this approach as the formation of fingers in the metal layer may serve only to concentrate the current at the tips of the fingers, and one could be effectively left with a reduced peripheral area.
8. Making holes in the central portion of the metal plate in order to provide internal peripheral edges to block the lateral flow of current. Some early claims were made that holes in the metal layer improved current density under the electrode. Presumably, it was believed that the holes blocked the current from flowing from the connector to the edge of the metal plate, forcing it to flow into the patient at the edges. It is the authors belief that such holes in the metal plate achieve little apart from further decreasing the area of the electrode and, if anything, increasing the current density at the edges. This impression appears to be confirmed by the work of Krasteva and Papazov (110), who investigated electrode structures with openings in the metal plate for skin breathing.

Figure 37. Medtronic's EDGE system electrode.

Figure 38. An electrostimulation electrode with cut-out metal plate in an effort to increase peripheral edge. The green sponge impregnated with gel has largely been removed to facilitate inspection of the underlying plate.

The author has suggested that the use of concave slits in the metal layer rather than circular holes may well have a favorable effect on current density distribution with the concave internal peripheral edges effectively blocking the lateral flow of current, forcing the trapped current to flow into the gel and, thus achieving a more uniform current density distribution over the surface of the conductive layer (115). Early work on the project with an industrial

partner appeared promising, but the work was never completed.

Resistive Electrodes. A TENS electrode system appears relatively simple and generally comprises a conductive plate, an ion-containing gel, a means of attachment to the skin, and a means of connection to the stimulators lead. Mannheimer and Lampe (42) pointed out, however, that of all the component parts of the overall TENS system, the electrode–skin interface has probably been the least understood and the most problematic. In addition to influencing the effectiveness of the treatment, poor electrode design can give rise to electrically, chemically, and mechanically induced skin irritation and trauma to the patient.

Initially, electrodes originally designed for ECG and other biosignal monitoring applications were used with TENS units, and some still are. Larger, more suitable electrode designs were eventually developed in order to reduce the current densities under the electrodes, to reduce skin irritation problems, and to increase stimulation comfort (116).

A large percentage of commercially available TENS electrodes are now molded from an elastomer (e.g., silicone rubber) or a plastic (e.g., ethylene vinyl acetate) and loaded with electrically conductive carbon black (Fig. 39) (3). Very few irritation or allergic reactions have been reported for conductive rubber electrodes as they do not generate the corrosion products often observed with metal electrodes (42). The great advantage of such electrodes is that they can be molded into almost any size or shape and a wide range of choice exists in the market. They can be made sufficiently thin to have high flexibility and, thus, are able to conform with body contours, making them suitable for a wide range of TENS applications.

Conductive rubber electrodes are often used in conjunction with an electrolyte gel and attached to the patient using elastic straps or custom-cut disks or patches of adhesive tape. Expanded polyester foam tends to give the most secure adhesion. However, as this backing is occlusive, the use of foam can give rise to skin irritation problems. Breathable, cloth-like fabrics allow the transmission of air and moisture and generally cause less skin irritation problems. Cloth-like materials tend to stretch, however (an advantage when it accommodates skin stretching due to movement), which can lead to the electrode working loose and making poor contact with the skin, possibly resulting in current density hot spots.

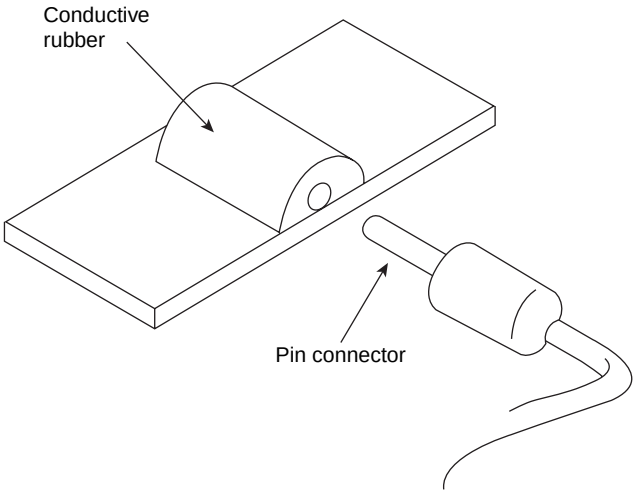


Figure 39. Carbon-filled silicone rubber electrode (3).

Wet gels can squeeze out from under parts of the electrode and give rise to increased current densities in other areas. The use of hydrogel minimizes this problem source (68). Conductive, adhesive pads of solid hydrogel help ensure firm electrical contact between the electrode and the skin, reduce the incidence of current density hotspots, and often simplify the design of the electrode. As a large surrounding disk of adhesive tape is not required, the electrode size can be reduced to the active electrode area. These solid gel pads can be, depending on the application, replaced, refreshed, or simply reused in various semi- or totally-reusable electrode systems. In some applications, the gel pads can be removed and the conductive rubber electrode cleaned and regelled with a fresh gel pad for further use. In other cases, the electrode can be intermittently reused, on the same patient, by rehydrating the gel pad. Such reusable electrodes are ideal for home-based patient use.

One disadvantage with such conductive rubber electrodes is that they are relatively resistive. More power is required to drive the stimulating current through the resistive electrodes into the body and achieve the desired stimulation. Therefore, some reduction in battery life may occur which is generally not a significant problem, however.

A more serious problem involves current density distribution under the resistive electrode. When current is applied through the conductive rubber (via a small metallic connector), it tends to flow into the skin immediately under the connector rather than laterally through the resistive electrode, thus giving rise to a current density hotspot under the connector, which effectively, is the opposite problem to that encountered when using highly conductive electrodes.

Efforts to overcome this problem include incorporating conductive elements in the rubber to more evenly to help spread the current over the entire interface surface. Some electrodes have a thin metallic layer coated onto the back of the conductive rubber, which appear to give rise to the most uniform current density profiles (68).

The growing home-based market has led to the great variety of low cost disposable and reusable electrodes that are generally based on solid adhesive gels. Some electrodes are made using conductive cloth-like materials, thin metallic foils, aluminized carbon-filled mylar, or wire strands. Electrical connection is generally made to these electrodes via alligator clips, snap fasteners, or pin connectors. Many of these hydrogel-based electrodes can be trimmed to the desired size or shape by simply cutting with a pair of scissors. The current density profiles under these electrodes will very much depend on the relative resistivities of the metal and gel layers as well as on the actual design.

Snap fastener designs resembling standard ECG electrodes and are available with hydrodel pads or sponge

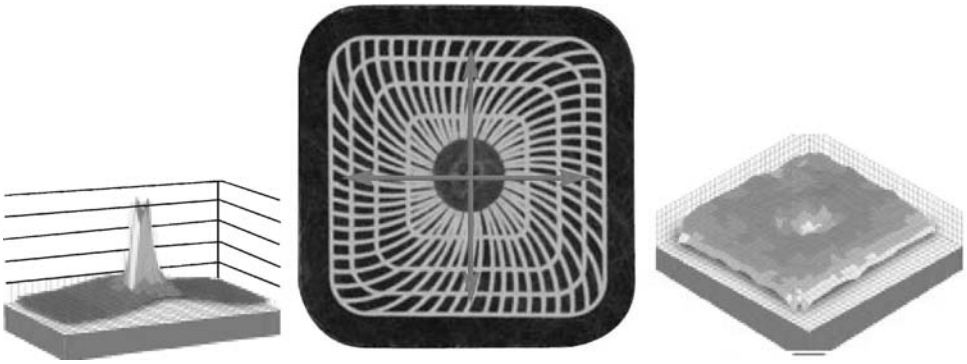


Figure 40. (a) Current density distribution under a conventional snap electrode. (b) Axelguaard's UltraStim snap electrode with current controlling grid. (c) Current density distribution under Axelguaard's UltraStim snap electrode. (Courtesy of Axelgaard Ltd.)

disks containing low chloride wet gels (to minimize skin irritation). These desists may require an additional current-dispersing element to ensure that the current spreads out beyond the immediate confines of the eyelet electrode (68). Axelgaard Ltds UltraStim Snap Electrodes feature a highly conductive grid pattern printed on to a conductive flexible layer and coated with a moderately conductive adhesive gel layer. The conductivity of the conductive pattern is controlled through the use of various grid designs with preselected line widths and spacing as well as thickness and ink compositions (117). The pattern is thus used to control and optimize the spread of electric current over the surface of the electrode with an intentional current drop off toward the edge of the electrode (Fig. 40).

It is interesting to note that considering current density under conductive and resistive electrodes leads to a similar optimal design. To improve conductive electrodes, one places a resistive layer between it and the patient. To improve a resistive electrode, one puts a conducting layer either behind or in front of the resistive layer. Such sandwich electrode designs appear promising for a range of electrostimulation applications.

Garment Electrodes. A range of researchers in the TENS, FES, and body-toning areas of electrostimulation are endeavoring to incorporate electrodes in to body hugging garments to enable the convenient and accurate application of a (large) number of electrodes to the body part to be stimulated. The use of a large number of electrodes can enable, for example, several muscle groups to be stimulated together or sequentially in a coordinated manner to achieve a more natural movement of a limb. Garments are already on the market that resemble tight-fitting cycling shorts and have integrated wires and connectors for the attachment of standard TENS (or similar) snap electrodes prior to application. Other, more challenging designs include the integration of reusable electrodes into a stretchable garment.

Implant Electrodes

Implantable monitors/stimulators and their electrodes are used, or are being developed, for a wide range of applications, including cardiac pacing and defibrillation, cochlear implants; urinary control, phrenic nerve stimulation for respiration control; functional electrical stimulation of limbs; vagal stimulation for control of epilepsy, spinal stimulation for chronic pain relief, deep-brain stimulation for Parkinsons disease or depression, bonehealing, and several visual neuroprostheses.

Implanted monitoring electrodes are used to more accurately pick up the desired signal while minimizing the contributions of extraneous signals. Implanted stimulation electrodes deliver the applied waveform more selectively to the targeted tissue, making the therapy more effective and, as the stimulation electrode is generally implanted away from cutaneous pain receptors and afferent nerve fibers, more comfortable for the patient. One significant drawback, however, is the greater potential for damage from improper electrode design, installation, and use.

The design of an implant electrode will depend greatly on the anatomical structure it is to be implanted against, into, or around. Electrodes can be or have been implanted in, on, or near a given muscle; in, on, or around a given nerve; in, on, or around a given bone; in, on, or around the spinal cord; and in or on the surface of the cerebellar cortex.

A review of all of these designs is beyond the scope of this chapter. The reader is referred to the appropriate chapters in this Encyclopedia.

To facilitate this overview of some of the key design possibilities, two main application areas will be concentrated on: muscular and neural electrodes. Muscular (especially cardiac) electrodes, using more traditional electrode fabrication, are presented in a separate section. Neural electrodes will be largely covered in the section on newer microelectrodes constructed using thin-film and similar techniques. These categories are very loose and a considerable degree of overlap obviously exists between applications and the various electrode fabrication techniques. Once again, the reader is referred to the appropriate chapters in this Encyclopedia for more detailed descriptions of electrodes and their fabrication for specific applications.

Cardiac electrodes are the most important example of muscular electrodes. As cardiac pacemakers and defibrillators have the longest and most successful track records as implantable devices, much of the science underpinning the newer (and future) implantable devices (muscular, neural, and other) has been developed by the cardiac implant pioneers. Key contributions were made in the areas of implant electrodes, biomaterials, and powersources, to

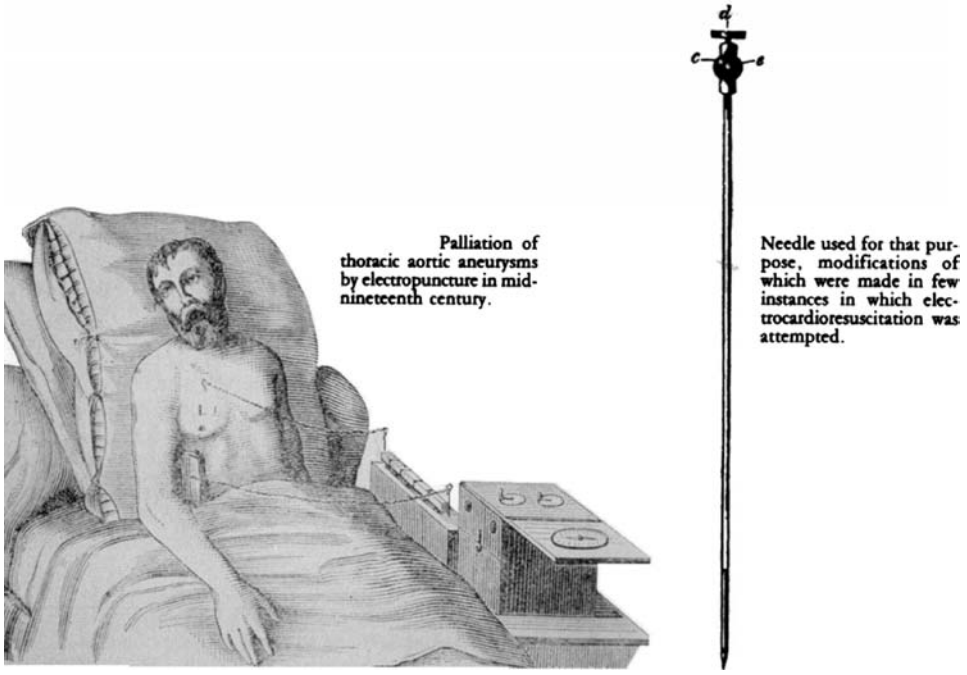


Figure 41. Kimer's electropuncture of the heart (103).

name but a few. The present review of electrodes is, therefore, largely based on cardiac electrodes. However, ideas can be gleaned from this area and, once suitably customized, applied to others with success. The advantageous aspects of biphasic impulses (discovered in the 1950s, arguably earlier) is a good example of something being rediscovered in a range of stimulation applications over the past few decades

Historical Background. Implant stimulation electrodes, or at least percutaneous stimulation electrodes, predate the earliest biosignal monitoring electrodes. In the early 1800s, there appeared a renewal of interest in acupuncture in Europe that had been introduced into Europe in the second half of the eighteenth century by Jesuit missionaries. In 1825, Sarlandière was the first to apply an electric (galvanic) current to thin metal needle electrodes (derived from acupuncture needles) thus creating electropuncture for the application of current to specific points on or in the body (98,118).

Electropuncture soon became the accepted method of stimulating muscles, nerves, or organs beneath the skin (83). Electropuncture of the heart was first attempted by Krimer in 1828 without recorded success (Fig. 41) (103). This technique was then abandoned for several decades. Meanwhile, W. Morton successfully introduced the use of ether as an anesthetic in 1846. Eventually, chloroform was found to be more suitable although cardiac arrest was a frequent complication of chloroform anesthesia in those early days. In 1871, Steiner overanesthetized horses, dogs, cats, and rabbits to produce cardiac arrest. He reported successfully applying an intermittent galvanic current to a percutaneous needle in the heart to evoke rhythmic contractions. Terms such as galvano and farado puncture soon started to appear in the literature (103).

In the early 1900s, cardio-stimulating drugs such as epinephrine were injected directly into the heart of sudden death victims by means of a large needle inserted through the chest wall to restore automatic activity. It was eventually established that one of the key factors in the occasional success of these intracardiac injection procedures was the actual puncture of the heart wall rather than the medication administered. Based on this observation, Hymen went on to build the first hand-cranked, spring-driven artificial pacemaker (119). He used transthoracic needle electrodes plunged into the atrium and even introduced the concept of using a biopolar needle arrangement as in having the two electrodes so close together that only a small pathway is concerned in the electric arc established by the heart muscle, an irritable point is produced (103).

In the 1950s, Lillehei, Weirich, and others pioneered the use of cardiac pacing for the management of heart block accidentally resulting from cardiac surgery and for other emergency cardiac treatment. Slender wire electrodes were implanted into the myocardium before closing the chest with the connecting leads thus exiting through the chest wall. Pacing impulses could then be delivered through these wires for a week or so until the heart healed. Once the heart had recovered, the electrodes were pulled out. Early versions of these electrodes consisted of silver-plated braided copper wires insulated with polyethylene or Teflon (103,120).

In 1958, Furman and Schwedel reported the first instance of transvenous pacing of the heart. They inserted a unipolar catheter electrode into the right ventricle of the patient through a superficial vein and paced the heart via the endocardial surface. The electrode used was a solid copper wire with a bare terminal tip (120). The electrode was withdrawn once the patients heart resumed its own idioventrcular rhythm. Although the cardiac pacing employed by Lillehei et al. and by Furman and Schwedel

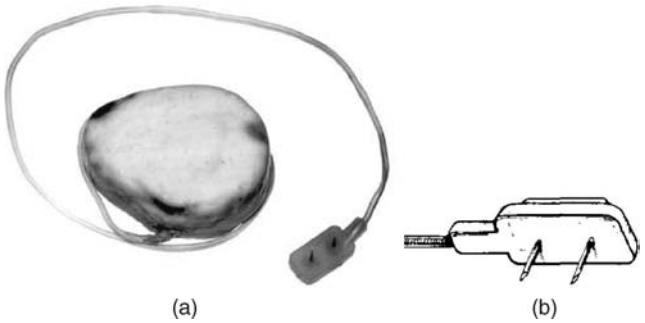


Figure 42. (a) Chardack, Gage, and Greatbatch's wholly implantable pacemaker and the Hunter–Roth bipolar intramyocardial electrode (83) (b) Diagram of an early Hunter–Roth lead with two bipolar myocardial pin electrodes (35).

was much better tolerated than the external stimulation of Zoll, perielectrode infection was a major drawback as was the transport of the external pacemaker. The development of an implantable pacemaker, therefore, became the goal of several groups around the world.

In 1958, Senning and Elmqvist successfully implanted the first Pacemaker without leads emerging from the patient's chest to invite infection. The implanted unit was powered by cells that were recharged from outside the body using a line-connected, vacuum-tube radio-frequency generator. The electrodes/leads used were stainless-steel wires. The second version of the unit failed due to a lead fracture one week following implantation. It was then decided to abandon pacemaker therapy for this patient until better leads were developed (120).

At that time, many of the electrodes used were unipolar. The active electrode tended to consist of the bared tip of an insulated wire implanted in the myocardium whereas the indifferent electrode was a similar wire implanted subcutaneously in the chest wall. Unfortunately, the stimulation threshold was observed to rise following implantation during longer-term pacing. This increase in threshold was thought to be due to the development of scar tissue around the active electrode. Hunter and Roth developed a bipolar electrode system in 1959. This electrode consisted of two rigid, 0.5 cm long, stainless-steel pins attached to a silicone rubber patch. The cathode–anode pins were positioned in to myocardial stab wounds surgically created for the purpose and the pad was then sutured to the epicardial surface (35). The lead wire was a Teflon-coated, multistrand stainless-steel wire with an outer sleeve made from silicone rubber tubing (121).

In 1960, Chardack, Gage, and Greatbatch successfully produced a wholly-implantable battery-powered pacemaker (Fig. 42a). Initially, they used a pair of multistrand stainless steel wires in a Teflon sleeve with the bare ends sutured to the myocardium (35). Other metallic formulations were tried, such as solid wire, silver wire, stainless steel, orthodontic gold, and platinum and its alloys (122).

They eventually adopted the Hunter–Roth intramyocardial electrode (Fig. 42a and b). Considerable surgery was required as the pacemaker had to be implanted into the abdomen and the electrodes were sutured to the heart wall. The bipolar electrode did, however, dispense with the need of a dispersive chest electrode and the associated pain it caused (103). Stimulation thresholds tended to stabilize at much lower levels with this electrode (120), which enabled successful pacing for many months.

Breakage of lead wires, due to metal fatigue, was a major concern. One of the main problem areas occurred at the point were the two metal components were welded together (121). Corrosion also occurred at the small-area stainless-steel anode, causing cessation of pacing within a few months.

Chardack et al. (123) devised a replacement for the Hunter–Roth electrode based on a continuous helical coil of platinum-iridium (Fig. 43). The electrode was simply a few turns of the coiled lead wire, exposed and extended to enable fibrous tissue to grow between the spirals and firmly anchor the electrode in place. The use of a helical coil greatly increased flexibility and decreased the number of fatigue failures, as did the use of one continuous wire (without a join) for both lead and electrode. The use of the same metal for lead and electrode also had the advantage of preventing corrosion from galvanic action. Additionally, platinum-iridium is more corrosion resistant than the metals used in many electrodes prior to Chardack's electrode.

The sutureless screw-in lead was later introduced by Hunter in 1973 (35). The screw-in electrode was simply rotated into the myocardium and did not require a stab wound or sutures for insertion. The electrode was effectively the means of attachment. As this corkscrew electrode tended not to dislodge, it dominated pacing for a long time and is still used today for many epimyocardial implants (Fig. 44).

A thoracotomy was required to attach many of the above electrodes to the heart, which complicated surgical procedure and resulted in a 10% early mortality (122). The first so-called modern pacemaker, which combined an implanted generator and a transvenous lead, was developed simultaneously in 1962 by Parsonnet and Lagergren (124,125). The endocardial catheter electrodes could be

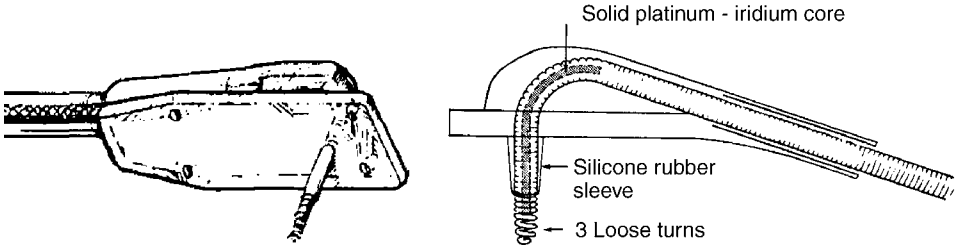


Figure 43. The "Chardack" electrode (a) (35) (b) Chardack et al. (123).

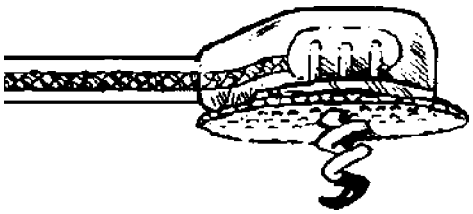


Figure 44. The corkscrew myocardial electrode (35).

installed under local anesthesia, and this approach virtually eliminated early mortality. As they did not require the opening of the chest cavity, the use of catheter leads opened the field of pacemaker implantation to non-surgeons in later decades.

To minimize the risk of venous perforation, the electrode leads were made flexible by winding bands of stainless steel around a core of textile fibers (120). The electrode was a small stainless-steel cylinder at the end of the catheter.

As time passed, the transvenous route progressively evolved over the myocardial approach, so much so that, at present, the transverse route is almost exclusively used for pacemaker implantation.

Cardiac pacing has been the earliest and most successful example of implanted electrodes and associated hardware. Many present and future developments in other implanted electrostimulation (and biosignal recording) areas are and will be based, to a large extent, on the pioneering work carried out in the pacing area.

Some Modern Electrode Designs. With the early transvenous leads, the stimulation threshold was observed to greatly increase if the electrode pulled away even slightly from the myocardium. A wide variety of active fixation devices was therefore invented. These devices included springs, deployable radiating needles, barbs, hooks, claws and screws designed to anchor the electrodes by actively penetrating the myocardium (35,126). The "Bisping" transvenous screw-in electrode is the most popular, as it allows the screw helix to be extended from the tip once the lead has been successfully threaded through the vein and located against the desired part of the heart (Fig. 45). It can be used as a combined anchor and electrode. The screw can be retracted allowing for an easier extraction of the lead, when necessary. (*Note:* A similar design of electrode is used for detecting the fetal electrocardiogram during labor. The intracutaneous needles are screwed in to the fetus' presenting scalp. Similar designs are also used in EEG monitoring.)

A wide variety of passive fixation devices were also invented. Various tines, flanges, and other soft, pliant projections were formed at the distal end of the lead, generally as an extension of the silicone or polyurethane lead insulation, and designed to passively and atraumatically wedge the electrode between endocardial structures such as trabeculae (Fig. 46). In some designs, the electrode has the form of a closed-loop helical coil that, when twisted clockwise, becomes lodged in the trabeculae (126).

Early electrodes had smooth metal surfaces. Techniques were then developed to roughen the surface in the hope of encouraging tissue in-growth, thus locking the electrode in place, minimizing mechanical irritation and excessive fibrous encapsulation, and ensuring low chronic stimulation thresholds. Studies found that porous electrodes did indeed achieve better fixation, thinner fibrous capsules, and stable thresholds.

A variety of porous electrode tips have been developed including totally porous structures such as CPIs meshed screen electrode and electrodes whose surfaces had been textured using a range of techniques (Fig. 47). Porous surfaces have been generated by coating metal surfaces with metallic granules, by sintering metal spheres to form a network of cavities, and by laser-drilling the surface of electrodes (126).

Not only does roughening improve electrode fixation and threshold stability, it has been found to have a very advantageous effect on interface impedance. From a stimulation point of view, one is keen to use a small-area electrode to increase current density at the small tip and

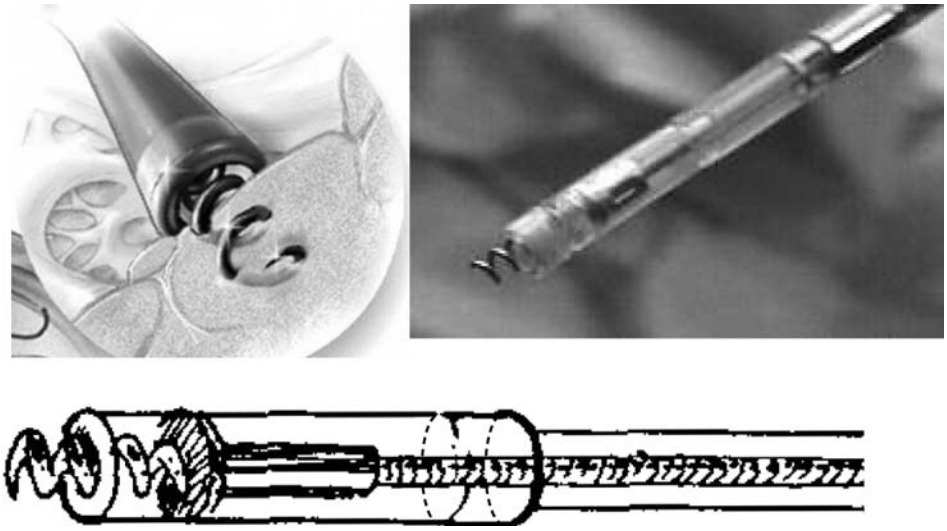


Figure 45. The "Bisping" transvenous screw-in lead with the helical screw electrode extended. [From S.S. Barold's The Third Decade of Cardiac Pacing (35).]

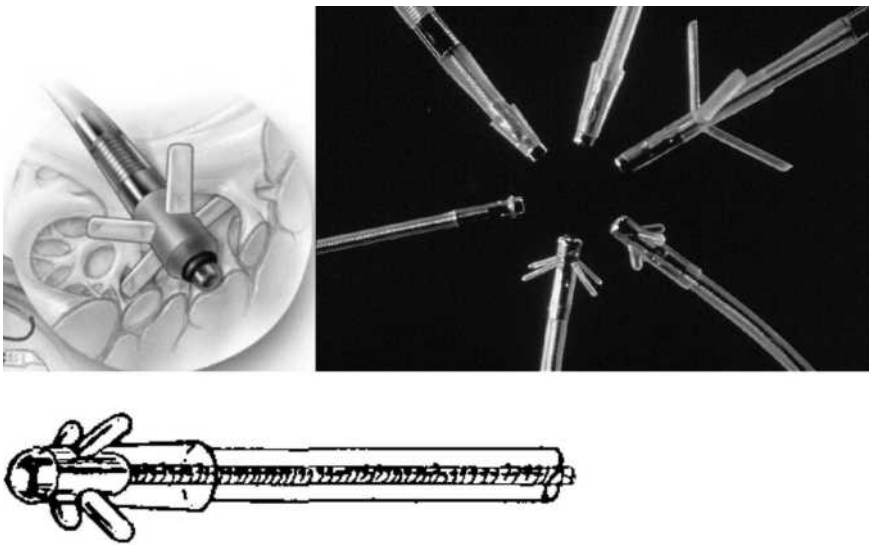


Figure 46. Transvenous lead with silicone rubber tines (35).

thus decrease stimulation threshold. A small-area electrode also has a high pacing impedance that can decrease current drain on the generator and thus prolong implant life (35). However, one would also like to have a low sensing impedance in order to avoid excessive attenuation of the cardiac signal. Two ways that exit these conflicting criteria can be optimized, modifying the electrode surface or modifying its design.

Roughening the surface of a small-area electrode increases its effective area without changing its geometric or outer envelope surface area (35). Many electrode systems have been developed that incorporate this concept—electrodes with terms such as activated, porous, and sintered in their company's description. These electrodes have been found to be effective in lowering the electrode interface impedance under small-signal sensing conditions. [Note: the electrode–electrolyte interface is very nonlinear and, hence, smaller under stimulation.] Unfortunately, the reduction in interface impedance has been erroneously interpreted as rendering the electrode nonpolarizable. As stated previously, the word polarization appears to be used in a rather vague manner and has been used as the explanation of, among other things, the nonlinearity of the interface impedance as well as its frequency- and time-dependence. The fact that the current or voltage response to a step in voltage or current is not a simple step has been attributed to polarization. The observed transient responses are merely due to the presence of the double layer capacitance (see Figs. 8, and 9). Roughening the surface of an electrode effectively increases the area of the interface and the value of C_{dl}, which in turn results in an increase in the response's time constant ($T = R_{CT}C_{dl}$). The observed response thus looks stretched out along the time axis. This flattened response has been mistaken for that of a purely resistive, nonpolarizable electrode. At any rate, roughening the surface of the electrode almost gives us the best of both worlds, a noble or inert electrode with a low interface impedance.

Another way of achieving a small simulation surface area (high current density) while ensuring a large-sensing surface area (low interface impedance) is to modify the design of the electrode.

The porous electrode of Amundson involved a hemispherical platinum screen that enclosed a ball of compacted 20 μm diameter platinum–iridium fibers (127). As electrolyte could penetrate this three-dimensional (3D) or multilayered electrode, the design resulted in a major increase in

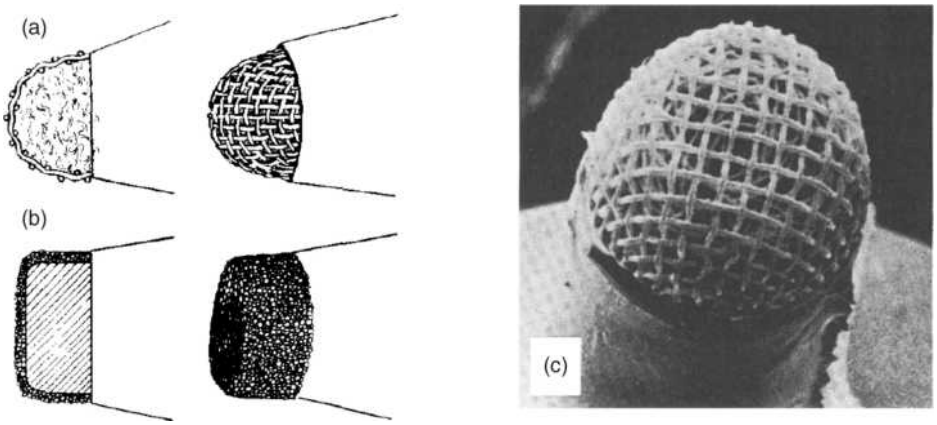


Figure 47. (a) Cross-section of a totally porous electrode (35). (b) Cross-section of a porous surface electrode (35). (c) Photo of totally porous electrode (126).

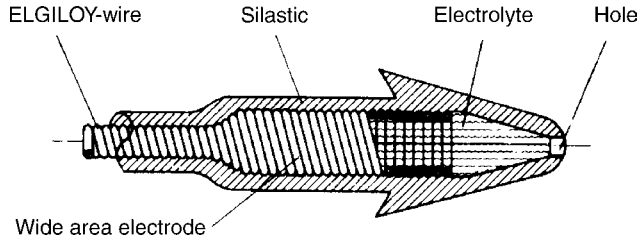


Figure 48. The differential current density (DCD) electrode (129).

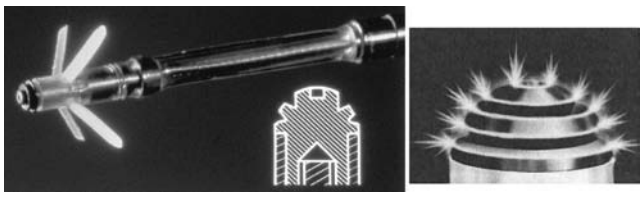


Figure 49. The target tip electrode. Microporous, plantinized plantinum electrode. The target appearance is due to shallow grooves separated by peaks. (Courtesy Medtronic, Inc.)

effective surface area as well as promoting tissue in-growth and long-term stability of thresholds. Lagergren et al. (128) introduced the birdcage design, which also exploited some of these features (126).

One interesting example of a modified design by Parsonnet et al. involved the use of an electrolyte-filled hollow electrode, called a differential current density (DCD) electrode (129). The actual stimulating electrode is the mouth of the electrolyte filled pore, which can be small to provide high current density at the point of contact with tissue (Fig. 48). The inside of the hollow electrode chamber has a large metallic surface (a helical coil forming a cylinder) and thus gives rise to a low electrode-electrolyte interface impedance.

Figure 48 appears to be an electrode design that could readily be customized and used in a wide range of monitoring or stimulation applications. The electrode–electrolyte interface is effectively recessed and protected from any disturbance, a further advantage to those already listed above.

Several other designs exist that aim to achieve a similar effect by manipulating the current density distribution around an electrode tip. Electrodes with complex shapes have irregular patterns of current density with localized hotspots at points of greatest curvature (126). It is possible to exploit these areas of high current density for stimulation purposes while the larger overall surface area gives rise to a low interface impedance (130). A hollow, ring-tipped electrode (effectively similar to the DCD electrode) has a large current density at its annular mouth while having a large electrode–electrolyte interface area. Such electrodes are reported to have better stimulation thresholds and sensing characteristics than hemispherical designs and have proved popular. Several manufactures have combined this ring-tip design with increased surface porosity (126). Other related designs include a dish-shaped electrode for edge-focusing of current (with laser-drilled pores for interface impedance reduction) and a grooved hemispherical platinum electrode coated with platinum black particles (target-tip electrode, Fig. 49) (126,130).

Steroid-eluting electrodes were introduced in 1983 in an effort to minimize the growth of connective tissue. The first-generation electrode was made of titanium, with a platinum-coated porous titanium surface (Fig. 50). The electrodes incorporated a silicone core that was impregnated with a small quantity an anti-inflammatory corticosteroid (34). Upon implant, the steroid is gradually eluted into the interface between the lead electrode and the endocardium, reducing the inflammation and fibrosis that would normally occur. Steroid-eluting leads are characterized by a lower long-term capture threshold. Similar improvements in capture thresholds have been achieved (131,132).

Most cardiac electrodes now involve the combination of drugs and complex surface structures at the macro and micro scales.

For newer applications, such as Cochlear implant electrodes, an array of electrodes is involved. In some such multielectrode applications, one may be interested not only in the current density profiles under the surfaces of the individual electrodes, but in the interplay between the electrical fields produced by the electrodes in the array in the hope of achieving more effective stimulation or more effectively imitate physiological stimulation. For example, the Clarion hi-focus electrode system of Advanced Bionics Corp. incorporates 16 electrodes in a flexible array that are designed to deliver improved focused stimulation to the auditory nerve (133).

Microelectrodes. Over the past few years, exciting developments have taken place in areas of biomedical

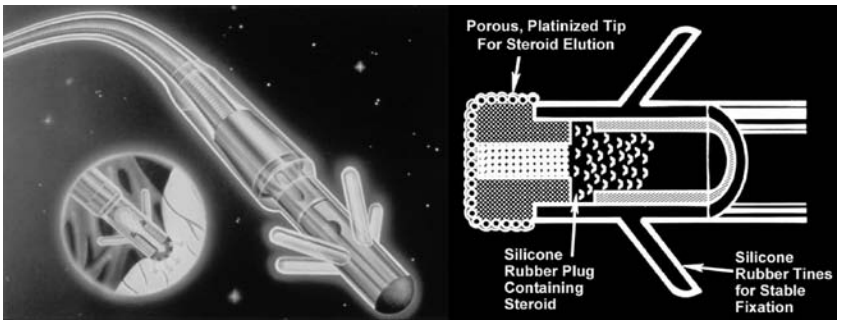


Figure 50. (a) Steroid-eluting electrode. (b) Cross-sectional diagram of an early design of steroid-eluting electrode. Behind the electrode is the silicone rubber plug compounded with steroid. (Courtesy Medtronic, Inc.)

engineering that involve implantable devices for the recording or stimulation of the nervous system.

In the previous section, we saw the success in commercializing pacemakers. Other implant devices that have also reached the patient in clinical routine practice or research settings include Cochlear implants to restore hearing; deep-brain stimulators to alleviate symptoms of Parkinson's Disease and depression; vagal nerve stimulators to minimize the effects of epilepsy; as well as FES systems to restore or improve function in the upper extremity, lower extremity, bladder and bowel, and respiratory system (134,135). Other areas of research that are likely to come to fruition within the next few years include various visual prostheses to restore functional vision in the profoundly blind and the exploration of the brain–computer interface (134,136,137).

Much of the early research in these areas started around the 1960s (135). Where possible and appropriate, surface and percutaneous electrodes were first used to establish the feasibility of the given recording/therapy. Early implant electrodes involved fine metallic wires or small disks placed near, in, on, or around the targeted muscle or nerve. The fabrication of these electrodes was time-consuming and the electrode properties were not very reproducible given the variations in areas, surfaces, inter-electrode distances, and so on, which was particularly a problem when several electrodes were to be used in an array. As the demands on human implantable diagnostic/stimulation devices increases, an increased need for a larger number of smaller-area electrodes with well-defined and reproducible surfaces and dimensions generally occors. Although, due to their high level of specificity, muscle-based electrodes will continue to be used, new electrode designs tend to concentrate more on direct nerve stimulation as this may provide more complete muscle recruitment and the same electrode may successfully recruit several muscles, thus reducing the number of electrode leads required (135). Electrodes are, therefore, needed that can interface electrically with the neural system at the micrometer scale (136).

For example, the goal for a high resolution retinal prosthesis is a 1000-electrode stimulating array in a 5×5 mm package (137). If this area of research is to be clinically successful and if the other areas are to continue to improve, microelectrodes must be (and are being) manufactured using the thin-film technologies associated with the IC circuit industry. Microfabrication involves either material deposition or removal. Either rigid silicon wafers or flexible polyimide substrates act as platforms for the microelectrodes and associated circuitry. The deposited films (for connectors, leads, electrodes, or insulation) are produced by electroplating, evaporation, and sputtering. The layers can be photo-patterned and etched to submicrometer resolutions and finally encapsulated in biomaterials such as diamond-like carbon, bioceramic, or a biocompatible polymer. Processes such as photolithography, reactive ion etching (RIE), CMOS processing, MEMS processing, focused ion beam patterning, and AFM lithography can be used to achieve the desired microelectrode design.

The benefits of a microfabrication approach include a high degree of reproducibility in physical, chemical, and electrical characteristics. Microfabrication is a high yield, low cost process once the design and processing sequence have been developed. Additionally, precise control of the spatial distribution of electrode sites exists, which may be of interest when seeking to optimally stimulate or record from a target site. A high packing density of electrode sites for a given implant volume is also readily achievable using photolithographic techniques. The possibility exits of incorporating the interface circuitry directly on the microsensor platform thus reducing the need for complex interconnections.

The widespread availability of silicon micromanufacturing techniques has enabled the fabrication of a range of silicon-based wedge- or needle-shaped electrodes to allow penetration of the nervous tissue. 3D arrays of such structures have been developed for insertion into, for example, the cortex to detect local potentials (134).

1D arrays of electrodes are fabricated using lithographic patterning and deposition of thin-film metal leads and electrodes onto not only silicon, but also glass and even flexible polyimide substrates (136). Much of the work on silicon-based microprobe fabrication has been pioneered at the Center for Integrated Sensors and Circuits at the University of Michigan.

A 3D electrode array can be fabricated by assembling a range of 1D probes (such as those shown in Fig. 51). As each probe has multiple recording sites along its length, the complete volume of the tissue under study can be assessed, giving rise to very dense sampling. The Michigan Probe has evolved a large number of single-shaft, multishaft, and 3-D-stacked microelectrode arrays (136).

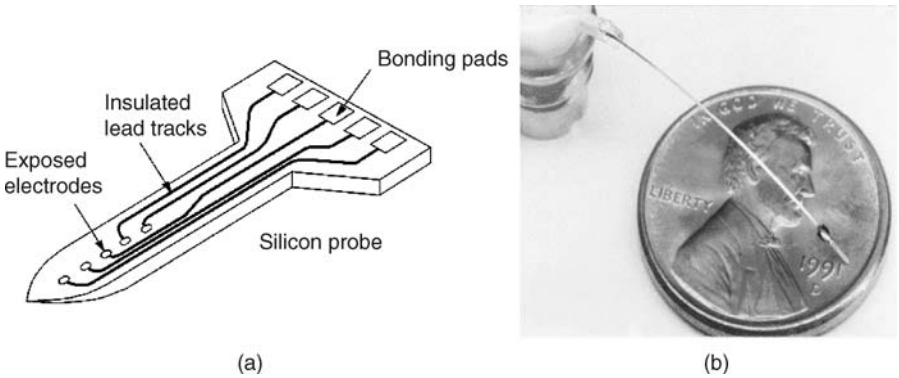


Figure 51. (a) Multielectrode silicon probe. [After Drake et al. (138).] (b) Michigan micromachined multielectrode probe for recording and stimulation of central nervous system.

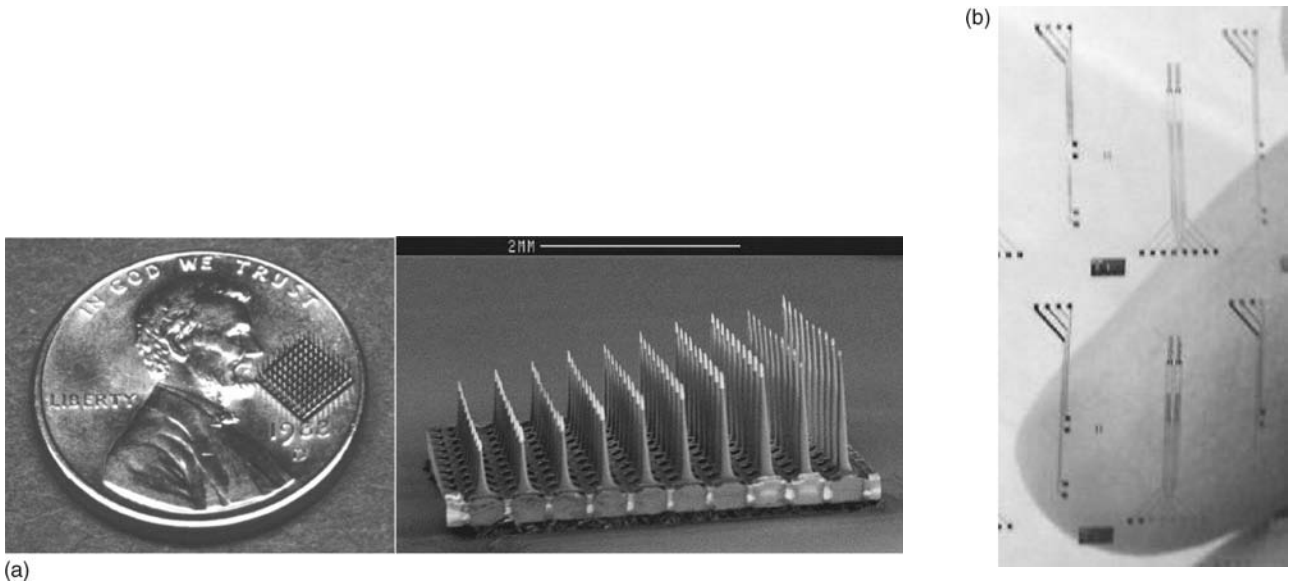


Figure 52. (a) Utah electrode array shown on a U.S. penny to convey size. (b) Modified Utah electrode array in which the length of the needles is uniformly graded (134).

Recent improvements in silicon microtechnology have made it possible to create not only planar microelectrodes but also penetrating brush electrode structures for *in vivo* measurements. In contrast to the University of Michigan's planar devices, the 2D and 3D cortical multimicroelectrode arrays developed at the University of Utah are fabricated out of a single solid block of silicon. Etching of the block results in a 10×10 array of needles, each 1.0–1.5 mm long, arranged on a 4.2×4.2 mm base. The metal and insulation layers are then applied, creating 35–75 μm long platinum recording tips (134,136,139) (Fig. 52a).

This design has the advantage of placing a relatively large number of recording sites in a compact volume of the cortex. However, with a single recording site on the end of each needle set at a fixed depth into the cortex, this version of the Utah array is classified as a 2D array as all the electrodes are in the same plane (140). The Utah probe can achieve high-density sampling by spacing many needles close together but does not have multiple sites along each shaft. When the length of the needles in such an array is graded (the array is said to be slanted, Fig. 52b) or the needles have some other distribution of lengths, these arrays are termed 3D as the electrode tips are no longer in the same plane. These designs are thought to give the better spatial selectivity (134,136).

Implanting such needle or brush electrode systems is obviously associated with damage of the tissue. Moreover, the stiffness of many systems may lead to damage of nervous tissue, especially if relative movement exists between the sharp needles and the delicate tissues. Breakage of the brittle needle is also a concern. Considerable efforts are therefore being directed at miniaturizing the width of the needles or at introducing more flexible materials.

For example, some versions of the Michigan Probe consist of four parallel, dagger-like probes connected to a micro-silicon ribbon cable. The ribbon cable is semiflexible and allows the probes to move up and down with the cortex as it pulses (139).

In the development of subretinal stimulating arrays using current silicon micromanufacturing techniques, it has been pointed out that a planar, rigid implant is likely to mechanically damage the compliant, spherical retina (137). Concerns have also been expressed regarding the use of penetrating microelectrodes, the relative micromotion between the array and the retina potentially provoking mechanical damage and a significant encapsulation response (134). The ideal retinal-stimulating electrode would therefore have the flexibility to match the curvature of the retina and the next generation of electrode arrays are likely to be constructed on flexible substrates.

Microelectrode arrays on flexible substrate have been demonstrated in a range of applications including the European project "Microcard", Si-Based Multifunctional Microsystem needle for Myocardial Ischemia Monitoring. Initially, work centered on silicon-based microprobes to monitor the electrical impedance of tissue, tissue temperature, pH, and local ionic concentrations of potassium, sodium, and calcium. These parameters were found to vary considerably when, for example, a heart undergoes an ischaemic phase, thus establishing the clinical value of the technique and device, (140).

In the course of the silicon probe development, it was foreseen that the brittle nature of silicon could make intact probe removal difficult. Additionally, the rigid needle could cause damage to the delicate tissues. The thrust of the project thus changed to the development of flexible, polymer-based probes.

Thin-film devices for the measurement of tissue impedance and ion concentrations were manufactured on flexible polyimide substrates (Fig. 53a) (141). Gold thin-film electrodes were deposited using an improved photolithography process for 1 μm resolution. Polyimide insulation

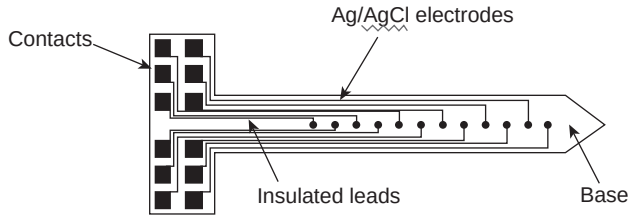


Figure 53. Microfabrication of sensors onto flexible substrates. 1D probe electrode array. [After Mastrototaro et al. (141).]

layers were spin-coated onto the PTFE surface after suitable conditioning, and they proved to be insulating and continuous.

Electrochemical characterization of the gold thin-film impedance electrodes showed them to possess high interface impedance. Pt and IrO oxide coatings were electrochemically applied to the gold thin-film surface and resulted in a drastic reduction in interface impedance for monitoring or stimulation applications (142).

Encircling neural electrodes may be of a cuff or spiral design. The term cuff electrodes applies to those devices that engulf the entire circumference of a nerve First model, which rather stiff, carried only one or two electrodes and they were made using a platinum foil electrodes that were located on the inside of a cylinder of silicone rubber, which was wrapped around a nerve (Fig. 54b). (136). It is generally recommended that the diameter of the cuff be 50% larger than the nerve diameter to avoid nerve compression and necrosis due to swelling and fibrous tissue ingrowth. Cuff electrodes do however have a long and successful track record in a range of FES applications (143).

The spiral electrode is a loose, open helix that is wound around the nerve (143). The open design can accommodate swelling and is very flexible. A version of this electrode is marketed by Cyberonics for use with their vagus nerve stimulator (Fig. 54b). New designs of nerve cuff electrodes seek to reshape the geometry of the nerve to more selectively stimulate or record from particular nerve fascicles. Efforts are also directed at controlling the electrical fields generated by the electrode arrays to better focus the stimulation (135).

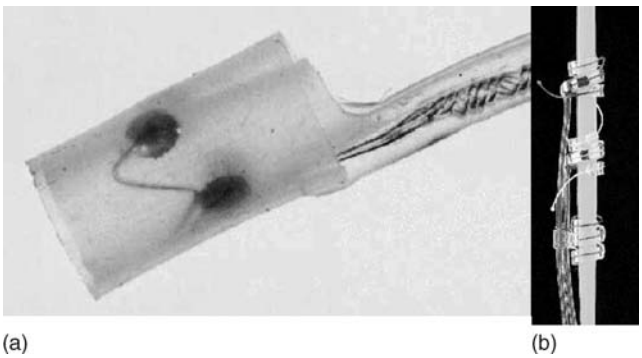


Figure 54. (a) Neural cuff electrode. (b) Spiral electrode (Cyberonics).

As part of a European project NEUROS, NIBEC developed a flexible thin-film-based stimulation and sensing cuff electrode for FES-related application. IrO, Pt, and Au electrodes were deposited onto a polyimide substrate. In order to facilitate implantation and ensure good contact between nerves, fascicles and electrode surface was self-curling. Polyimide resin with a thermal expansion coefficient differing from that of the polyimide substrate was chosen so that the curing process gives rise to a residual stress and curl in the device. The diameter of the electrode cylinder could be made less than 1 mm.

Diamond-like carbon (DLC) encapsulation was deposited onto the device using a plasma-enhanced chemical vapor deposition (PECVD) process. Adhesion to the polyimide substrate was found to be satisfactory following the addition of a silane adhesion layer at the interface (144,145).

With the aid of micofabrication techniques, one can control the area and properties of the electrodes and greatly decrease them in size. However, as electrode area decreases, the interface impedances increase with resultant difficulties in making accurate measurement. The key to success in this case is in the choice of electrode design, material, and electrode surface topography.

A similar concept to Chardack's differential current density pacemaker electrode was suggested for use in thin-film electrodes. The metal electrode is housed within a hollow chamber (Fig. 55). The chamber is filled with electrolyte and has a small aperture to enable electrical contact with tissues (146). As the metal–electrolyte interface is relatively large, the interface impedance is relatively small. The interface is also protected from mechanical disturbance (similar to the floating electrode) and, hence, should suffer from less artifact. As the small aperture determines the area of contact with the tissue, the effective stimulation or recording area is very small.

Other 3D designs with etched meshes should be assessed for their potentially larger interfacial areas.

Once again, surface roughness is an important factor in decreasing interface impedance and possibly in helping anchor the electrode in position. Rough-surfaced electrodes must be used with caution, depending on the application, in case the surface causes damage to the surrounding tissues. Certain materials and the electrode fabrication processes involved may well result in favorable macro-, micro-, and nano- surface features. Presently, investigators are studying modifications to the electrode surface using such things as nanotubes. Nanotechnology offers much promise for

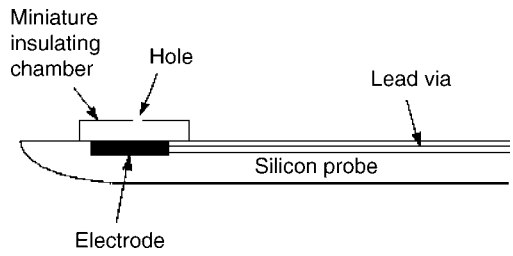


Figure 55. Thin-film differential current density electrode. [After Prohaska et al. (146).]

new sensor devices, particularly in the biomedical sector. Not only do individual nanotubes offer the possibility of using them as ultrafine needles for *in vivo* probing at the cellular level, but surfaces can be created with optimal distributions of clusters of nanotubes to maximize performance.

ELECTRODE STANDARDS

The Association for the Advancement of Medical Instrumentation (AAMI) produce a range of labeling, electrical, and other performance requirements for manufacturers and users to help ensure acceptable levels or product safety and efficacy. Some of the key electrode-related standards are briefly reviewed below.

Standards For Biosignal Monitoring Electrodes

Standards for Disposable ECG Electrodes. ANSI/AAMI EC 12 (2000)

Introduction. In an effort to minimize ECG recording problems associated with the performance of electrodes coupled to a standard ECG monitor or electrocardiograph, AAMI has proposed a series of simple bench tests designed to assess pregelled, disposable ECG electrodes.

Although originally conceived to assess disposable ECG electrodes, these standards are widely used to assess other biosignal monitoring electrodes, which is a consequence of the lack of other widely accepted standards for these monitoring applications and to the general applicability of the ECG standards to the other applications.

Although AAMI also lays down stipulations for electrode labeling, adhesion testing, and soon, only the electrical performance requirements are reviewed here.

AC Impedance. The average value of 10 Hz impedance for at least 12 electrode pairs connected gel-to-gel, at a level of impressed current not exceeding 100 μA peak-to-peak, shall not exceed 2 kΩ. None of the individual pair impedances shall exceed 3 kΩ.

Low impedance electrodes are desirable to avoid signal attenuation and distortion and to minimize 50/60 Hz interference pickup. High electrode impedances can also give rise to serious burns when the ECG electrodes are used in the presence of electrosurgery or defibrillator discharges. (147).

The impedance of the skin's outer layer, the stratum corneum, is many times larger than that of the metal/electrolyte interface, and hence, the former is of key concern when endeavoring to ensure good electrode performance. The skin preparation technique, the extent of diaphoresis, and the ability of the electrode gel in penetrating and reducing the skin impedance are generally more important than the electrode–electrolyte interface impedance.

The Standards Committee decided that the electrode gel-to-gel impedances should be significantly less than the expected impedance of clean, dry skin to ensure a minimal contribution by the electrode itself to the overall impedance (147). In the UBTL tests carried out on behalf of the Standards Committee, it was found that the mean 10 Hz impedance of a standard pair of ECG electrodes on unabraded skin was of the order of 100 kΩ. The AAMI committee chose 2000 kΩ as a reasonable limit for 10 Hz gel-to-gel impedance to ensure that the electrodes did not contribute significantly to the overall impedance nor to power dissipation in the presence of defibrillation overload and electrosurgery currents.

As the electrode-gel interface impedance is nonlinear and decreases with applied signal amplitude, the standard stipulates that the level of impressed current must not exceed 0.1 mA peak-to-peak when carrying out the test.

In the UBTL tests, it was found that the impedance as measured on abraded skin correlated well (99%) with the impedance measured with the electrodes connected gel-to-gel, whereas the impedance measured with electrodes applied to clean, dry skin correlates very poorly (47%) with the gel-to-gel measurements. Obviously, the bench test simply evaluates the ac impedance performance of the electrode-gel interface and will, therefore, not accurately predict or represent the clinical performance of an electrode on intact skin. For example, cases of electrodes that performed poorly as per the AAMI bench test exist, yet which proved very satisfactory *in vivo*. Conversely, some of the best electrodes according to the bench tests performed relatively badly in vivo (148).

DC Offset Voltage. After a 1 min stabilization period, a pair of electrodes connected gel-to-gel, shall not exhibit an offset voltage greater than 100 mV.

Ideally, the potentials of both electrodes used to monitor a biosignal should be identical and, thus, cancel each other out. Slight differences in the gels and metals used, however, result in an offset voltage. The potentials of the skin sites further complicate the recording, especially as these latter potentials (and their amplified difference) tend to be much larger. If the overall electrode-skin potential difference is larger than 300 mV, the amplifier may saturate and the biosignal will not be observed.

The UBTL report studied the correlation between gel-to-gel and electrode-skin offset voltages and found that gel-to-gel offsets were in the order of 2.5 times smaller than those recorded *in vivo* for the same electrodes on a patient's skin. As the maximum allowable *in vivo* dc offset should be less than 300 mV, the Committee decided that the limit for gel-to-gel dc offset should therefore be less than 300/2.5 mV (i.e., 100 mV).

Some reviewers of the standard argued that the limit should be reduced to 10 mV as this would help minimize motion artifact problems. The Committee rejected this suggestion, pointing out that no clear evidence exists that links high gel-to-gel offset voltages with motion artifact (largely caused by skin deformation).

Offset Instability and Internal Noise. After a 1 min. stabilization period, a pair of electrodes connected gel-to-gel, shall not generate a voltage greater than 150 $\mu V_{p\text{-}p}$ in the passband (first-order frequency response) of the 0.15–100 Hz for a period of 5 min following the stabilization period.

This standard is concerned with the problem of baseline wander, which introduces a low frequency component into the monitored biosignal making accurate diagnosis difficult. The American College of Cardiology's Task Force on the Quality of Electrocardiographic Records judged that drift rates less than 400 $\mu V \cdot s^{-1}$, although not highly rated, were not considered unacceptable.

The UBTL report detailed several experimental limitations that prohibited their detailed study of *in vivo* dc offset drift. Consequently, no correlational analysis was carried out between dc offset drift measurements made with electrodes applied to human skin and those joined gel-to-gel. They, however, decided to use the factor of 2.5 they had observed between clinical and bench test result for dc offsets, given that the measurement techniques are fundamentally similar. A limit of 150 $\mu V \cdot s^{-1}$ was therefore arrived at by dividing the 400 $\mu V \cdot s^{-1}$ baseline drift rating by a factor of 2.5. As the test circuit used in the bench test differentiates the offset voltage, the offset instability requirement is specified in μV rather than $\mu V \cdot s^{-1}$.

The Committee was contacted and asked to decrease the limit from 150 μV to 40 μV p-p in order to be in line with the AAMI standard "Cardiac monitors, heart rate meters and alarms (EC13)". The working group agreed that this requirement could be made more stringent but refused to decrease the limit to 40 $\mu V_{p\text{-}p}$. This requirement is under study and may well be altered.

This calculation involved in reaching the 150 $\mu V \cdot s^{-1}$ limit implies that skin potential fluctuations are only 2.5 times larger than those of the electrode–gel interface, which is most unlikely, and problems developing from drifting electrolyte/skin potentials will depend on skin preparation, electrode design, and electrode gel rather on than the electrode–gel interface characteristics per se (64).

Defibrillation Overload Recovery. Five seconds after each of four capacitor discharges, the absolute value of polarization potential of a pair of electrodes connected gel-to-gel shall not exceed 100 mV. Also during the 30 s interval following each polarization potential measurement, the rate of change of the residual polarization potential shall be no greater than $\pm 1\,mV \cdot s^{-1}$.

It is important that a clinician, having defibrillated a patient, be able to see a meaningful ECG within 5–10 s in order to judge the efficacy of the delivered impulse and to decide if another is required. The offset voltage across the electrode-skin interfaces, which drastically increased as a result of the defibrillation impulse, must therefore return to below 300 mV within 5 s following the discharge. Once again, using the 2.5 factor between bench test and *in vivo* potentials, this requirement translates to a gel-to-gel bench test offset voltage under 100 mV within 5 s of applying an overload of 2 mC (representing the worst possible situation encountered *in vivo* where the defibrillator paddles are placed in immediate contact with the ECG electrodes). Electrodes made of stainless steel, for example, tend to acquire offset voltages of several hundred mV for minutes and, consequently, no ECG trace is observable on the monitor (68).

Following the initial 5 s the EGC must not only be visible on the monitor but must also be recognizable and clinically useful. Hence, the stipulation that the offset voltage should not drift with time by more than $\pm 1\,mV \cdot s^{-1}$.

The UBTL results indicate good correlation exists between the results of this bench tests and animal tests, particularly at the higher recovery voltages encountered with non-Ag/AgCl electrodes.

Although only a very low percentage of ECG electrodes are, in fact, subjected to defibrillation impulses *in vivo*, the AAMI committee decided after some deliberation to insist that all ECG electrodes meet the proposed standard as it is impossible to guarantee that a given electrode would not be used in an emergency defibrillation situation.

Bias Current Tolerance. The observed dc voltage offset change across a pair of electrodes connected gel-to-gel shall not exceed 100 mV when the electrode pair is subjected to a continuous 200 nA dc current over the period recommended by the manufacturer for the clinical use of the electrodes. In no case shall this period be less than 8 h.

When a dc current passes through the metal-gel interface of an electrode, the electrode potential deviates from its equilibrium value and the electrode is said to be polarized. If the current is maintained indefinitely, the reactants become depleted causing the electrode potential to deviate further, possibly exceeding the limit allowable at the input of the ECG recording device.

Although most modern ECG recorders pass less than 10 nA of bias current through the electrodes, some older models can have bias currents as high as 1000 nA. A number of cardiac monitor manufacturers use dc bias currents to sense high electrode impedances to warn of disconnected leads or poorly affixed electrodes. The standard for cardiac monitors permits input bias currents of up to 200 nA. UBTL, therefore, adopted the 200 nA limit on the dc input bias current suggested for cardiac monitors for the tests. The ability of an electrode to cope with this value of bias current must therefore be demonstrated by not exceeding the AAMI dc offset requirement of 100 mV over the time period recommended by the manufacturer for the clinical use of the electrodes.

The 200 nA current level is generally well-tolerated by Ag/AgCl electrodes. Stainless-steel electrodes rapidly fail this test even at 10 nA with major increases in electrode potential.

Discussion. The AAMI standards bench tests are currently the only widely accepted electrode standard tests in use. The tests are simple and inexpensive to set up and have been widely embraced by manufacturers and users for production quality control purposes. One must bear in mind, however, that these tests evaluate only the electrode-gel interface and that they do not include the more important properties of the gel–skin interface. Assessment of the clinical performance of electrode impedance using the proposed bench tests is only relevant if the skin has been suitably abraded. Skin abrasion is not widely used by the clinical community and, hence, the relevance of at least some of the standard tests to the clinical situation is open to question.

Especially several decades ago, fulfillment of the AAMI requirements was commonly quoted as a guarantee

of the high *in vivo* electrical performance of an electrode. An electrode with, for example, a dc offset of 1 mV was widely believed by customers to be a much better electrode than one with an offset of 5 mV. This naivety appears to be on the wane, however, and manufacturers and customers are shifting toward low cost electrodes that score less highly in the AAMI tests but are good enough for a given application.

The author once supplied a leading company with dry metal-loaded polymer electrodes. The company connected the electrodes together and tested them as per the AAMI standards (for pregelled electrodes). Perfect electrical performances were measured given that what was effectively being assessed was metal-to-metal contact. Direct current offsets of 0 mV were obtain. Once the dry electrodes were applied to a patient's skin, a less than favorable result was obtained.

The attitude to adopt, therefore, when interpreting AAMI standard bench tests results for pregelled, disposable ECG electrodes is that electrodes that meet the AAMI standards have a tendency rather than a certainty to perform well *in vivo*. Electrodes that perform better as per the bench tests do not necessarily perform better *in vivo*. They are a useful set of tests nonetheless.

The ANSI/AAMI standard tests were conceived such that the test apparatus needed can be readily assembled by an electrode manufacturer. However, one can can buy a convenient-to-use, custom-built electrode tester (as per AAMI standards) called the Xtratek electrode tester ET65A (Direct Design Corporation, Lenexa, Kansas.) (Fig. 56).

Electrocardiograph surface electrode testers also exist for the *in vivo* testing of the quality of (*1*) the design ECG electrodes, (*2*) the application of the electrodes, and (*3*) the skin preparation technique used.

The electrode tester generally measures the ac impedance and dc offset of the electrode-patient system. These measurements can be used, for example, in stress testing to decide if the skin sites have been sufficiently well prepared (i.e., contact impedances are low enough) to proceed with the clinical procedure. They can also be used to detect the presence of loose cables or bad contacts.

Figure 56. Early version of the Xtratek electrode tester ET65A. (Direct Design Corporation; Lenexa, Kansas.)

Standards for Stimulation Electrodes

Although not covered in this article, the following standards exist that stipulate minimum labeling, safety, and performance requirements for the given stimulators. The rationale for the standards is also presented.

- Transcutaneous electrical nerve stimulators ANSI/AAMI NS4.
- Implantable spinal cord stimulators ANSI/AAMI NS14.
- Implantable peripheral nerve stimulators ANSI/AAMI NS15.

Standards for Automatic External Defibrillators and Remote-Control Defibrillators. ANSI/AAMI DF 80 (2003)

AC Small Signal Impedance. The 10 Hz impedance for any of at least 12 electrode pairs connected gel-to-gel, at a level of impressed current not exceeding 100 μA peak-to-peak, shall not exceed 3 kΩ. The impedance at 30 kHz shall be less than 5 Ω. The rationale for this requirement is based on the performance criteria in ANSI/AAMI EC 12 for disposable ECG electrodes. Interestingly, the permissible gel-to-gel 10 Hz impedance for large-area defibrillation pads is higher than that allowed for small-area ECG electrodes. The gel-to-gel impedance measured at 30 kHz will be largely that of the gel pads as the interface impedances at this frequency will be almost zero.

AC Large Signal Impedance. The impedance of an electrode pair connected gel-to-gel, in series with a 50 Ω load and measured at the maximum rated energy of the defibrillator shall not exceed 3 Ω. A value of 50 Ω is thought to represent the typical (rather low) *in vivo* transthoracic impedance between the electrodes. One wants the delivered energy to be dissipated in the patient's chest and not in the electrodes where the wasted energy may give rise to skin burns. The above requirement is therefore thought to provide a reasonable limit on the impedance contributed to the overall impedance by the electrode pair during defibrillation (<6%).

Combined Offset Instability and Internal Noise. A pair of electrodes connected gel-to-gel shall generate, after a 1 min stabilization period, a voltage no greater than 100 μV peak-to-peak in the pass band of 0.5–40 Hz, for a period of 5 min following the stabilization period. The rationale for this requirement is based on the performance criteria in ANSI/AAMI EC 12 for Disposable ECG electrodes. The frequency range used is more limited in recognition that the cardiac monitor bandwidth is more appropriate in this application.

Defibrillation Recovery. The potential of a pair of gel-to-gel electrodes in series with a 50 Ω resistor and subjected to three shocks at 360 J or maximum energy at 1 min intervals shall not exceed 400 mV at 4 s and 300 mV at 60 s after the last shock delivery. The rationale for this requirement is largely based on the performance criteria in ANSI/AAMI EC 12 for Disposable ECG electrodes. An actual

defibrillation impulse is applied instead of that from a simulation circuit. The offset voltage across the simulated electrode-patient load must return to below 400 mV within 4 s following the discharge (slightly different values, 300 mV and 5 s, are used in ANSI/AAMI EC 12). As the patient's chest is represented by the 50 Ω resistor, no need exists for the 2.5 factor used in ANSI/AAMI EC 12 to correlate bench test and *in vivo* results.

DC Offset Voltage. A pair of electrodes connected gel-to-gel shall, after a 1 min stabilization period, exhibit an offset voltage no greater than 100 mV. The rationale for this requirement is based on the performance criteria in ANSI/AAMI EC 12 for disposable ECG electrodes.

Universal-Function Electrodes. With conventional defibrillators, it has been customary to use separate pregelled ECG electrodes for monitoring and defibrillator paddle electrodes for defibrillation. The monitoring electrodes are not capable of effectively delivering a defibrillation shock, and the paddle electrodes have only limited monitoring capability. For recent applications, particularly automatic external defibrillation, it is very desirable to use self-adhesive pregelled disposable combination electrodes that perform well in the dual monitoring and defibrillation functions. These electrodes may also be used for delivery of transcutaneous pacing. Hence, combination electrodes may become preferred for defibrillation, and it is appropriate in a standard for defibrillators to consider their use and to outline a few requirements for them.

If the electrodes are designed and intended for use in multiple modes (i.e., monitoring, defibrillation, and pacing) the electrode shall meet all (of the above) requirements after 60 min of pacing at the maximum current output and maximum pacing rate through a pair of gel-to-gel electrodes in series with a 50 Ω resistor.

No general performance standards exist for combination pacing/defibrillation/monitoring electrodes, the (above) requirements define the basic minimum controls necessary to ensure safe and reliable operation.

Standards for Electrosurgical Devices. ANSI/AAMI HF 18 (2001)

Introduction. Although AAMI lays down stipulations for the testing of a range of parameters, only the key electrical performance requirements for the dispersive electrodes are reviewed below.

Maximum Safe Temperature Rise. The maximum patient tissue temperature rise shall not exceed 6 °C when the dispersive electrode carries a current of 700 mA under the test conditions below, unless the device is labeled in accordance with 4.1.4.2 (i.e., for use on infants). For devices labeled for use on infants, the maximum patient tissue temperature rise shall not exceed 6 °C when the dispersive electrode carries a current of 500 mA under the test conditions stipulated in the standard. In monopolar electrosurgical procedures, the dispersive electrode must be able to reliably conduct the required surgical current without generating a significant rise in skin temperature. It is widely accepted that the maximum safe skin temperature for short-term and long-term exposure is 45 °C, as normal resting skin temperature varies between 29° and 33 °C. Electrodes must not generate skin temperature increases approaching 12 °C. A 6 °C increase in temperature is therefore thought to represent an acceptable upper limit.

The temperature measurement method must have an overall accuracy of better than 0.5 °C and a spatial resolution of at least one sample per square centimeter of the electrode thermal pattern. The thermal pattern must include the area extending 1 cm beyond the geometry of the electrode under test. This degree of special resolution is stipulated as electrosurgical burns may be confined to very small areas and these must be detected. As current tends to flow to the edge of the electrode and spread out further in the skin, the test requires that the surround area of skin is also scanned.

The electrode under test is to carry a current from an electrosurgical generator of 700 mA_{rms} for 60 s, unless the device is labeled in accordance with 4.1.4.2, in which case the test current may be 500 mA. A current of 700 mA applied for 60 s yields a heating factor of 30 A^2 s. [Heating Factor $= I^2t$ (A^2s).] This value is far in excess of the maximum likely current and duration for a TUR (transurethral resection) procedure. A more realistic heating factor is less than 10 and, hence, the stipulated testing procedure is very conservative.

These tests must be conducted on human volunteers or on a suitably structured surrogate medium. When human volunteers are used, the tester must include a variety of body types in the sample group rather than concentrate on a single body type (thin, average, or thick layers of subcutaneous body fat). If surrogate media are used, the tester must demonstrate that the media are electrically and thermally similar to human volunteers. Human volunteer subjects are the reference standard. Current density distribution under an electrode depends on a wide range of factors, including the electrical properties of the skin and underlying tissues, hence, the need to test a given electrode on a wide range of individuals. The use of a surrogate material, even pig skin, which is commonly used, will not necessarily replicate with sufficient accuracy the clinical performance of the electrode. If a surrogate medium is used, the tester must demonstrate the equivalence of the test medium to human tissue. It is the Committee's view that no adequate surrogate medium has yet been suggested or used that has all of the properties of human tissue for the purpose of determining electrode performance.

Nessler et al. (149) point out that the above experiments are laborious, time-consuming, and expensive to perform. They have developed a new test device, swaroTEST, which includes a surrogate electronic skin, which, they claim, simulates the relevant electrical features of human skin and thus can replace the required volunteer experiments (Fig. 57). The device consists of a 3D resistor network representing the electric features of the skin and muscle tissue, and a temperature-sensing array (one transistor for each cm^2) to measure the resultant temperature increase after a standardized current load (700 mA hf current during 60 s, proposed in the relevant AAMI HF-18 standard). The authors claim that a comparison of results obtained with their device and those with thermo camera images of

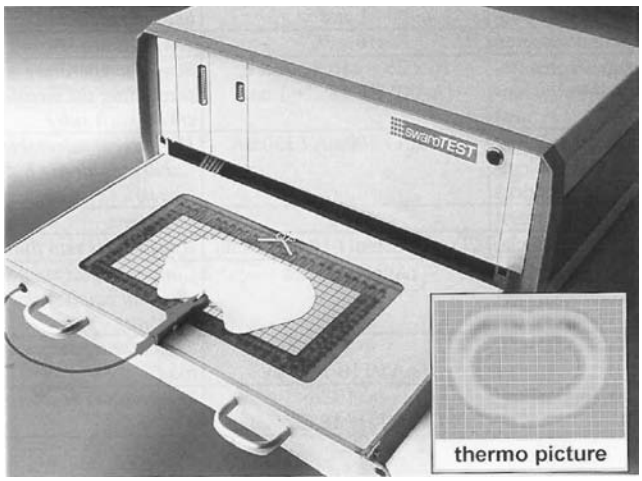


Figure 57. Swaro Test device with a measuring board electronic skin to simulate the electrical properties of human skin. It is hoped that this device will replace the human volunteer experiments required for ANSI.AAMI HF 18.

volunteer experiments correspond sufficiently well to justify the acceptance of their test device as a surrogate medium.

Electrode Contact Impedance. The electrode contact impedance must be low enough that the dispersive electrode represents the preferred current pathway, thus avoiding skin burns at alternative pathways. For conductive electrodes, the maximum electrode contact impedance shall not exceed 75 Ω over the frequency range of 200 kHz–5 MHz when measured as described on a human subject. The frequency range of 200kHz–5 MHz encompasses the frequency ranges of existing generators. As electrode-tissue impedance increases as applied current decreases, the committee decided on an impedance measuring current of 200 mA as it represents the lower limit of average currents reported for TUR procedures. Under these conditions, a maximum contact impedance value of 75 Ω was judged an acceptable for the conductive electrodes.

For capacitive electrodes, the minimum capacitance shall be no less than 4 nF (0.004 μF) when measured as described. In this case, electrode contact impedance is measured by placing the capacitively coupled dispersive electrode under test on a rigid metal plate larger than the electrode contact area. The test current and frequencies are the same as those specified for conductive electrodes. Their impedance characteristics are described in terms of capacitance as their impedances vary as the inverse of the frequency. The majority of capacitive electrodes that have been found to be clinically acceptable typically have a capacitance value of 4 nF, hence the minimum acceptable capacitance value specified by the Committee.

SUMMARY

With external biosignal monitoring electrodes, difficult challenges exist in the exciting new area of personalized health. Such electrodes must form part of the patient's (or health-conscious citizen's) clothing and must continue to work, day after day, wash after wash, without gelling or preparation of any kind, without suffering from motion artifacts, and without causing skin irritation, which is no mean achievement.

An old monitoring problem still remains to be adequately conquered. A convenient and rapid method of applying many high performance electrodes to the head of a patient for EEG measurement awaits invention. The problem (and that of ECG ambulatory monitoring above) can be side-stepped to some extent by finding new electrode positions (montages or leads) that avoid the most problematic skin sites, hairy head in EEG and muscle and flabby areas in ECG.

For external stimulation, exciting new areas include public access defibrillation. The electrodes and their application to the victim must be almost literally fool-proof, given the seriousness of the possible consequences for all concerned. The electrodes must work after having been stored in the most inhospitable locations and possibly under extreme temperature fluctuations, for example, in the trunk of a car in the desert. In the more mainstream areas of cardiac pacing and defibrillation and electrosurgery, the optimal distribution of current density under the electrodes remains a goal still to be achieved. The solution to this problem offers the hope of decreased electrode areas and the design of truly multifunction pads.

The integration of electrodes into garments for FES and body toning is a relatively new area with considerable possibilities.

At present, implant electrodes and associated technologies already offer amazing potential for the deaf, lame, and even the blind. The development of multimicroelectrode arrays and waveforms that can help optimally shape the electrical fields to facilitate more effective and natural stimulation is a thrilling prospect. The interface properties of the microelectrodes will require further research so as to offset the potentially high interface impedances. Ideas already exploited in cardiac pacing, for example, may prove rewarding when adapted for these areas.

It is hard to overstate the potential of research being undertaken in the area of brain–machine interface. We live in exciting times.

BIBLIOGRAPHY

Cited References

1. Gatzke RD. In: Miller HA, Harrison DC, editors. The electrode: A measurement systems viewpoint. Biomedical Electrode Technology. New York: Academic Press; 1974.
2. Janz GJ, Ives DJG. Silver-silver chloride electrodes. Ann NY Acad Sci 1968;148:210–221.
3. Webster JG. Medical Instrumentation: Application and Design. 3rd ed. New York: John Wiley & Sons; 1998.
4. Cheney M, Isaacson D, Newell JC. Electrical impedance tomography. SIAM Rev 1999;41:85–101.
5. McAdams ET, Jossinet J, Lackermeier A, Risacher F. Factors affecting the electrode-gel-skin interface impedance in electrical impedance tomography. Med Biol Eng Comput 1996;34(6):397–408.

6. Singh S, Singh J. Transdermal drug delivery by passive diffusion and iontophoresis: A review. Med Res Rev 1993;13(5):569–621.
7. Bard AJ, Faulkner LR. Electrochemical Methods. New York: John Wiley & Sons; 1980.
8. Almasi JJ, Hart MW, Schmitt OH, Watanabe Y.Bioelectrode voltage offset time profiles and their impact on ECG measurement standards. Can Med Biol Eng Conf 4th, Winnipeg, Manitoba, Canada, 1972.
9. McAdams ET. Effect of surface topography on the electrode-electrolyte interface impedance, Part 1: The high frequency, small signal interface impedance. Surface Topogr 1989;2: 107–122.
10. Fricke H. The theory of electrolytic polarization. Philos Mag 1932;7:310–318.
11. Cole KS, Curtis HJ. Transverse electric impedance of squid giant axon. J Gen Physiol 1938;22:3764.
12. Sluyters-Rechbach M, Sluyters JH. Sine wave methods in the study of electrode processes. In: Bard AJ, editor. Electroanalytical Chemistry, vol. 4. New York: Marcel Dekker; 1970. pp 1–128.
13. De Levie R. The influence of surface roughness of solid electrodes on electrochemical measurements. Electrochim Acta 1965;10:113–130.
14. de Levie R. On the impedance of electrodes with rough interfaces. J Electroanal Chem 1989;261:1–9.
15. Maritan A, Toigo F. On skewed arc plots of impedance of electrodes with an irreversible electrode process. Electrochim. Acta 1990;35:141–145.
16. McAdams ET. Effect of surface topography on the electrode-electrolyte interface impedance, Part 2: The low frequency ($F<1$ Hz), small signal interface impedance. Surface Topogr 1989;2:223–232.
17. Bergveld P. Med Biol Eng Comput 1976;14:479–482.
18. Brummer SB, Robblee LS, Hambrecht FT. Criteria for selecting electrodes for electrical stimulation: Theoretical and practical considerations. Ann NY Acad Sci USA 1983;405: 159–171.
19. Brummer SB, Turner MJ. Electrical stimulation of the nervous system: The principle of safe charge injection with noble metal electrodes. Bioelectrochem Bioenerget 1975; 2:13–25.
20. Dymond AM. IEEE Trans BME 1976;23:274–280.
21. Lilly JC, Hughes JR, Alvord EC, Galkin TW. Brief, noninjurious electric waveform for stimulation of the brain. Science 1955;121:468–469.
22. Weinman J. Biphasic stimulation and electrical properties of metal electrodes. J Appl Physiol 1965;20:787–790.
23. Fischler H, Schwan HP. Polarisation impedance of pacemaker electrodes: *In vitro* simulating practical operation. Med Biol Eng Comput 1981;19:579–588.
24. Schwan HP. Electrode polarization impedance and measurements in biological materials. Ann NY Acad Sci USA 1968;148:191–209.
25. Schwan HP. Alternating current electrode polarisation. Biophysik 1966;3:181–201.
26. Simpson RW, Berberian JG, Schwan HP. Nonlinear AC and DC polarization of platinum electrodes. IEEE Trans Biomed Eng 1980;27:166–171.
27. Jaron D, Briller SA, Schwan HP, Geselowitz DB. Nonlinearity of cardiac pacemaker electrodes. IEEE Trans Biomed Eng 1969;16:132–138.
28. Onaral B, Schwan HP. Linear and non-linear properties of platinum electrode polarization: Part 1.Frequency dependence at very low frequencies. Med Biol Eng Comput 1982;20:299–306.
29. McAdams ET, Henry P, Anderson JMcC, Jossinet J. Optimal electrolytic chloriding of silver ink electrodes for use in electrical impedance tomography. Clin Phys Physiol Meas 1992;13(Suppl 1):19–23.
30. McAdams ET, Jossinet J. The importance of electrode-skin impedance in high resolution electrocardiography. Automedica 1991;13:187–208.
31. McAdams ET, Jossinet J. A Physical Interpretation of Schwan's limit voltage of linearity. Med Biol Eng Comput 1994; March: 126–130.
32. McAdams ET. Jossinet J. DC nonlinearity of the solid electrode-electrolyte interface impedance. Inn Tech Biol Med 1991;12:329–343.
33. McAdams ET, Jossinet J. A physical interpretation of Schwan's limit current of linearity. Ann Biomed Eng 1992; 20:307–319.
34. K Stokes. Cardiac pacing electrodes. Proc IEEE 1996;84(3): 457–467.
35. Stokes K. Implantable pacing lead technology. IEEE Eng Med Biol 1990;9(2):43–49.
36. Williams DF. The Williams Dictionary of Biomaterials. Liverpool University Press; 1999.
37. Geddes LA. Electrodes and the Measurement of Bioelectric Events. New York: John Wiley & Sons; 1972.
38. Crenner F, Angel F, Ringwald C. Ag/AgCl electrode assembly for thin smooth muscle electromyography. Med Biol Eng Comput 1989;27:346–356.
39. Kingma YJ, Lenhart J, Bowes KL, Chambers MM, Durdle NG. Improved Ag/AgCl pressure electrodes. Med Biol Eng Comput 1983;21:351–357.
40. Geddes LA, Baker LE, Moore AG. Optimum electrolytic chloriding of silver electrodes. Med Biol Eng 1969;7:49–56.
41. Heath R. Tin-stannous chloride electrode element. U.S. Patent 4,852,585, 1989.
42. Mannheimer JS. Lampe GN. Clinical Transcutaneous Electrical Nerve Stimulation. Philadelphia, F.A. Davis, PA; 1987.
43. Prausnitz MR, Bose VG, Langer R, Weaver JC. Electroporation of mammalian skin: A mechanism to enhance transdermal drug delivery. Proc Natl Acad Sci USA 1993;90:10504–20508.
44. Brown L, Langer R. Transdermal derlivery of drugs. Ann Rev Med 1988;39:221–229.
45. Rosendal T. Further studies on the conducting properties of human skin to direct and alternating current. Acta Physiol Scand 1945;8:183–202.
46. Rosendal T. Concluding studies on the conducting properties of human skin to alternating current. Acta Physiol Scan 1945;9:39–49.
47. Salter DC. A study of some electrical properties of normal and pathological skin in vivo. Ph.D. dissertation, University of Oxford. Oxford (UK): 1980.
48. Klingman AM. Skin permeability: Dermatologic aspects of transdermal drug delivery. Am Heart J 1984;108(1):200–207.
49. Reilly JP. Electrical Stimulation and Electropathology. Cambridge, UK: Cambridge University Press; 1992.
50. Chien YW. Transdermal controlled-release drug administration.In: Swarbrick J. editor Novel Drug Delivery Systems. New York: Marcel Dekker Inc.; 1982. p 149.
51. Edelberg R. Electrical properties of the skin. In: Elden HR, editor. A Treatise of the Skin. New York: John Wiley & Sons; 1971.
52. Yamamoto Y, Yamamoto T. Dispersion and correlation of the parameters for skin impedance. Med Biol Eng Comput 1978;16:592–594.

53. Chien YW. Development of transdermal drug delivery systems. Drug Develop Industr Pharm 1987;13(4&5):589–651.
54. Rothman S. Electrical behavior. In:Physiology and Biochemistry of the Skin. Chicago (IL): The University of Chicago Press; 1956. p 9–25.
55. Lawler JC, Davis MJ, Griffith EC. Electrical characteristics of the skin. J Invest Dermatol 1960; 301–308.
56. Rosell J, Colominas J, Riu P, Pallas-Areny R, Webster JG. Skin impedance from 1 Hz to 1 MHz. IEEE Trans Biomed Eng 1980;35:649–651.
57. Almasi JJ, Schmitt OH. Systemic and random variations of ECG electrode system impedance. Ann N Y Acad Sci 1970;170:509–519.
58. Grimnes S. Dielectric breakdown of human skin in vivo. Med Biol Eng Comput 1983;21:379–381.
59. Schmitt OH, Almasi JJ. Electrode impedance and voltage offset as they affect efficacy and accuracy of VCG and ECG measurements. Proc. XIth International Vectorcardiography Symposium, New York, 1970; 245–253.
60. Yamamoto T, Yamamoto Y. Analysis for the change of skin impedance. Med Biol Eng Comput 1977;15:219–227.
61. Searle A, Kirkup L. A direct comparison of wet, dry and insulating bioelectric recording electrodes. Physiol Meas 2000;21:271–283.
62. McAdams ET, Lackermeier A, Woolfson ET, Moss GP, McCafferty DF, In vivo ac impedance monitoring of percutaneous drug delivery. Proc. 9th Int. Conf. on BioImpedance, Heidelberg, Germany, 1995: 344–347.
63. De Talhouet H, Webster JG. The origin of skin-stretch-caused motion artefacts under electrodes. Physiol Meas 17:81–93.
64. Tam HW, Webster JG. Minimizing motion artifact by skin abrasion. IEEE Trans Biomed Eng 1977; BME 24:134–140.
65. Zinc R, Distortion and interference in the measurement of electrical signals from the skin (ECG, EMG, EEG). Innovation and Technology in Biology and Medicine, 12, special issue. 1991; 1: 46–59.
66. McLaughlin J, McAdams ET, Anderson JMcC. Novel dry electrode ECG sensor system. 16th Annual Int Conf IEEE Eng Med Biol Soc Baltimore (MD), Nov. 1994:804.
67. Jossinet J, McAdams ET. Skin Impedance. Innovation and technology in biology and medicine, 12, special issue. 1991;1:21–31.
68. Carim HM. Bioelectrodes. In: Webster JG. editor Encyclopaedia of Medical Devices and Instrumentation. New York: Wiley & Sons; 1988. p 195–226.
69. Oh SY, Leung L, Bommannan D, Guy RH, Potts RO. Effect of current, ionic strength and temperature on the electrical properties of skin. J Controlled Release 1993;27:115–125.
70. Olson WH, Schmincke DR, Henley BL. Time and frequency dependence of disposable ECG electrode-skin impedance. Med Instrum 1979;13:269–272.
71. McAdams ET. Surface biomedical electrode technology. Int Med Device Diagnost Ind 1990; 44–48.
72. McAdams ET, McLaughlin JA, Anderson J McC. Multi-electrode systems for electrical impedance tomography. Physiol Meas 1994;15:A101–A106.
73. McAdams ET, McLaughlin J, Brown BN, McArdle F. In: London HD, editor. The NIBEC EIT harness, Clinical and Physiological Applications of Electrical Impedance Tomography. Chapt 8, UCL Press; 1993. p 85–92.
74. McAdams ET, Jossinet J. Hydrogel electrodes in bio-signal recording. Proceedings of the 12th Annual International Conference of the IEEE, Philadelphia, PA: Engineering in Medicine and Biology Society; 1990: 1490–1491.
75. McAdams ET, Lackermeier A, Jossinet J. AC impedance of the hydrogel-skin interface. 16th Annual Int. Conf IEEE Eng in Med and Biol Soc Baltimore (MD), 1994: 870–871.
76. Carim HM, Hawkinson RW. EKG electrode electrolyte-skin AC impedance studies. Proc. 4th Ann Conf IEEE Eng Med Biol Soc 1982:503–504.
77. Yamamoto T, Yamamoto Y. Electrical properties of the epidermal stratum corneum. Med Biol Eng 1976;14:151–158.
78. Geddes LA. A. Historical perspectives 2: The electrocardiograph. In: Bronzino JD, editor. The Biomedical Engineering Handbook. Boca Raton FL: CRC Press; 1995; p 788–798.
79. Waller AD. A demonstration on man of electromotive changes accompanying the heart's beat. J Physiol 1887; 8:229–234.
80. Waller AD. On the electromotive changes connected with the beat of the mammalian heart, and of the human heart in particular. Phil Trans R Soc London Ser B 1989;180:169–194.
81. Waller AD. Introductory address on the electromotive properties of the human heart. Brit Med J 1888;2:751–754.
82. Barker LF:. Electrocardiography and phonocardiography: A collective review. Bull Johns Hopkins Hosp 1910;21:358–359.
83. Rowbottom ME, Susskind C. In: Electricity and Medicine: History of their Interaction. San Francisco (CA): San Francisco Press; 1984.
84. Barron SL. The development of the electrocardiograph in Great Britain. Br Med J 1950;1:720–725.
85. Lewes D. Multipoint electrocardiography without skin preparation. Lancet 1965;2:17–18.
86. Wolferth CC, Wood FC. The electrocardiographic diagnosis of coronary occlusion by the use of chest leads. Am J Med Sci 1932;183:30–35.
87. Barnes AR, et al. Standardization of precordial leads. Am Heart J 1938;15:235–239.
88. Burch GE, DePasquale NP: A History of Electrocardiography with a New Introduction by Joel D Howell, 2nd ed. San Francisco, CA: Jeremy Norman; 1990.
89. Ungerleider HE. A new precordial electrode. Am Heart J 1939;18:94.
90. Welch W. Self-retaining electrocardiographic electrode. JAMA 1951;147:1042.
91. Jasper HH, Carmichael L. Electrical potentials from the intact human. Science 1935;81:51–53.
92. Khan A, Greatbatch W. Physiologic electrodes. In: Ray CD. editor. Medical Engineering. Chicago, IL: Year Book Medical Publishers; 1974.
93. Manley AG, Medical electrode. US patent 3,977,392, 1976.
94. K Krug, Marecki NM. Porous and other medical and pressure sensitive adhesives. Adhes Age 1983;26(12):19–23.
95. Hymes AC. Monitoring and stimulating electrode. U.S. Patent 4,274,420, June 23, 1981.
96. Dempsey GJ, McAdams ET, McLaughlin J, Anderson JMcC. NIBEC cardiac mapping harness. 14th Annual Int. Conf. IEEE Eng. In Med and Biol Soc Paris, France, Nov 1992: 2702–2703.
97. Lymberis A. Research and Development of Smart Wearable Health Applications: The Challenge Ahead, Wearable eHealth Systems for Personalised Health Management, Studies in Health Technology and Informatics 108. Lymberis A, de Rossi D, editors, IOS Press; 2004.
98. Axisa F, Schmitt PM, Gehin C, Delhomme G, McAdams E, Dittmar A. Flexible technologies and smart clothing for citizen medicine, home healthcare and disease prevention. IEEE Trans Inform Technol Biomed 2005;9(3): 325–336.
99. Adams G. An Essay on Electricity. London; 1785.

100. Aldini G. Account of Late Improvements in Galvanism. London; 1803.
101. Duchenne GBA. In: De l'Électrisation Localisée et de son Application á la Physiologie, á la Pathologie et à la Thérapeutique. 1855.
102. Duchenne GBA. In: Baillière JB et al., editors. Mécanisme de la Physionomie Humaine. 1876.
103. Schechter DC. In: Exploring the Origins of Electrical Cardiac Stimulation. Medtronic; 1983.
104. Robinson AJ, Snyder-Mackler L. Clinical Electrophysiology: Electrotherapy, Electrophysiologic Testing. Baltimore (MD): Williams and Wilkins; 1995.
105. Low J, Reed A, Electrotherapy Explained: Principles and Practice. Oxford: Butterworth–Heinmann Ltd; 1994.
106. Stankevich BA. 4% of professional liability claims involve electromedicine equipment. Mod Health Care 1980;10(12): 74–76.
107. Pearce JA. The thermal performance of electrosurgical dispersive electrodes. Ph.D. dissertation. Purdue University, West Lafayette (IN); 1980.
108. Wiley JD, Webster JG. Analysis and control of the current distribution under circular dispersive electrodes. IEEE Trans Biomed Eng 1982;29:381–385.
109. Caruso PM, Pearce JA, DeWitt DP. Temperature and current density distributions at electrosurgical dispersive electrode sites. Proc 7th N Engl Bioeng Conf., Troy, New York, March 22–23, 1979: 373–376.
110. V Krasteva, Papazov S. Estimation of current density distribution under electrodes for external defibrillation. BioMedical Engineering Online, 2002; 1:7. Available http://www.biomedical-engineering-online.com/content/1/1/7.
111. Y Kim, Schimpf PH. Electrical behavior of defibrillation and pacing electrodes. Proc IEEE 1996;84(3):446–456.
112. Kim Y, Fahy JB, Tupper B. Optimal electrode designs for electrosurgery, defibrillation, and external cardiac pacing IEEE Trans Biomed Eng 1986;33:845–853.
113. Netherly SG, Carim HM. Biomedical electrode with lossy dielectric properties. US pat 5,836,942, 1998.
114. Ferrari RK. X-ray transmissive transcutaneous stimulating electrode. US pat 5,571,165, 1996.
115. McAdams ET, Andrews P. Biomedical electrodes and biomedical electrodes for electrostimulation. US pat 2003, 134,545, 2003.
116. Szeto AYJ. Pain relief from transcutaneous electrical nerve stimulation (TENS). In: Webster JG. ed. Encyclopedia of Medical Devices and Instrumentation. New York: John Wiley & Sons; 1988. p 2203–2220.
117. AXELGAARD J. Reverse current controlling electrode. US pat 2004,158,305, 2004.
118. Sarlandière. "Mémoires sur l'électropuncture considérée comme moyen nouveau de traiter efficacement la goutte, les rhumatismes et les affections nerveuses. Paris, 1825.
119. Hyman AS. Resuscitation of the stopped heart by intracardial therapy. Arch Intern Med 1932;50:283.
120. Mittal T. Pacemakers – A journey through the years. Ind J Thorac Cardiovasc Surg 2005;21:236–249.
121. Myers GH, Parsonnet V. Pacemaker electrodes In: Myers GH, Engineering in the Heart and Blood. New York: Wiley-Interscience; 1969.
122. Greatbatch W, Holmes CF. History of implantable devices. IEEE Eng Med Biol 1991; Sept: 36–49.
123. Chardack WM, Gage AA, Greatbatch W. Correction of complete heart block by a self-contained and subcutaneously implanted pacemaker. J Thorac Cardiovasc Surg 1961;42: 418.
124. Lagergren H, Johansson L. Intracardiac stimulation for complete heart block. Acta Chir Scand 1963;125:562–566.
125. Parsonnet V, Zucker IR, Asa MM.Preliminary investigation of the development of a permanent implantable pacemaker utilizing an intracardiac dipolar electrode. Clin Res 1962; 10:391.
126. Timmis G. The electrobiology and engineering of pacemaker leads. In: Saksena S, Goldschlager N. eds. Electrical Therapy for Cardiac Arrhythmias. New York: Saunders W.B. Co.; 1990.
127. Admundson DC, McArthur W, Mosharrafa M. The porous endocardial electrode. PACE 1979;2:40–50.
128. Lagergren H, Edhag O, Wahlberg I. A low threshold nondislocating endocardial electrode. J Thorac Cardiovasc Surg 1976;72:259.
129. Lewin G, Myers GH, Parsonnet V, Zucker IR. A non-polarizing electrode for physiological stimulation. Trans Am Soc Artific Intern Organs 1967;13:345.
130. Ellenbogen KA, Wood MA. Cardiac Pacing and ICDs. New York: Blackwell Science Inc; 2002.
131. Mond H, Stokes KB. The electrode-tissue interface: The revolutionary role of steroid elution. PACE 1991;15:95–107.
132. Mond HG, Stokes KB. The steroid-eluting electrode: A 10-year experience. Pacing Clin Electrophysiol 1996 Jul; 19(7):1016–1020.
133. Lenarz T, Battmer R-D, Goldring JE, Neuburger J, Kuzma J, Reuter G. New electrode concepts (Modiolus–Hugging Electrodes). Adv Otorhinolaryngol Basel Karger 2000;57:347–353.
134. Maynard EM. Visual prostheses. Annu Rev Biomed Eng 2001;3:145–168.
135. Peckham PH, Knutson JS. Functional electrical stimulation for neuromuscular applications. Annu Rev Biomed Eng 2005;7:327–360.
136. Rutten WLC. Selective electrical interfaces with the nervous system. Annu Rev Biomed Eng 2002;4:407–452.
137. Weiland JD, Liu W, Humayun MS. Retinal prosthesis. Annu Rev Biomed Eng 2005;7:361–401.
138. Drake KL, Wise KD, Farraye J, Anderson DJ, BeMent SL. Performance of planar multisite microprodes in recording extracellular single-unit intracortical activity. IEEE Trans Biomed Eng 1988;35:719–732.
139. Schwartz AB. Cortical neural prosthetics. Annu Rev Neurosci 2004;27:487–507.
140. Aguiló J. Microprobe multisensor for graft viability monitoring during organ preservation and transplantation. 2nd Annual International IEEE-EMB Special Topic Conference on Microtechnologies in Medicine & Biology, Madison, WI. February 2002; 15–20.
141. Mastrototaro JJ, Massoud HZ, Pilkington TC, Ideker RE. Rigid and flexible thin-film multielectrode assays for transmural cardiac recording. IEEE Trans Biomed Eng 1992; 39:271–279.
142. Linquette-Mailley SC, Hyland M, Mailley P, McLaughlin J, McAdams ET. Electrochemical and structural characterisation of electrodeposited iridium oxide thin film electrodes applied to neurostimulating electrical signal. Mater Sci Eng 2002;21:167–175.
143. Mortimer JT, Bhadra N. Peripheral nerve and muscle stimulation. In: Horch KW, Dhillon GS, editors. Neuroprosthetics: Theory and Practice (Series on Bioengineering & Biomedical Engineering), vol. 2. New York: World Scientific; 2004. p 638–744.
144. Hyland M, McLaughlin J, Zhou DM, McAdams E. Surface modification of thin film gold electrodes for improved in vivo performance. Analyst 1996;121:705–709.
145. Rieger R, Taylor J, Comi E, Donaldson N, Russold M, Mahony CMO, McLaughlin JA, McAdams E, Demosthenous A, Jarvis JC. Experimental determination of compound A-P

direction and propagation velocity from multi-electrode nerve cuffs. Med Biol Eng Comput Phys 2004;26:531–534.
146. Prohaska OJ, Olcaytug P, Pfundner P, Dragaun H. Thin film multiple electrode probes: Possibilities and limitations. IEEE Trans Biomed Eng 1986;33:223–229.
147. Schoenberg AG, Klingler DR, Baker CD, Worth NP, Booth HE, Lyon PC. Final report:Development of test methods for disposable ECG Electrodes. UBTL Technical Report No. 1605–005, Salt Lake City (UT); 1979.
148. Hollander JI, ECG-Electrodes.Report No. 83.336, MFI-TNO, Utrecht, The Netherlands, 1983.
149. Nessler N, Reischer W, Salchner M. Electronic skin replaces volunteer experiments. Measure Sci Rev 2003;3(2):71–74.

Reading List

Bell GH, Knox AC, Small AJ. Electrocardiography electrolytes. Br Heart J 1939;1:229–236.
Geddes LA, Baker LE. Principles of Applied Biomedical Instrumentation, 3rd edition. New York: John Wiley & Sons; 1989.
Licht S. History of electrotherapy. In: Licht S. ed. Therapeutic Electricity and Ultraviolet Radiation. New Haven, CT: Elizabeth Licht Pub; 1959. p 1–69.

See also DEFIBRILLATORS; ELECTROCARDIOGRAPHY, COMPUTERS IN; ELECTROENCEPHALOGRAPHY; ELECTROSURGICAL UNIT (ESU); FUNCTIONAL ELECTRICAL STIMULATION; TRANSCUTANEOUS ELECTRICAL NERVE STIMULATION (TENS).

BIOFEEDBACK

JOHN C. ARENA
VA Medical Center and Medical Collage of Georgia
TRISHUL DEVINENI
Conemaugh Health System
EDWARD J. MCGOWAN
E.J. McGowan & Associates

INTRODUCTION

Biofeedback is a term that first arose in the 1960s for a methodology that uses instrumentation to record the physiological responses of organisms and then in real time give information about those physiological responses back to the organism. It is presumed that by getting such timely feedback about physiological responding, the organism will learn, through a trial and error basis, how to control the desired physiological response.

The most concise definition of biofeedback is probably that of Olton and Noonberg (1), who characterized it as, "any technique that increases the ability of a person to control voluntarily physiological activities by providing information about those activities" (p. 4). In practice, the process of clinical biofeedback training involves the use of a machine (usually a computer-based system in contemporary applications), which allows a therapist to monitor the patient's bodily responses (most commonly surface muscle tension or surface skin temperature). Information concerning the patient's physiological responses are then relayed back to the patient, generally either through an auditory modality (a tone that goes higher or lower depending on, say, electrical activity of the target muscles increasing or decreasing) and/or a visual modality (now usually a computer screen where, e.g., surface skin temperature is sampled and then graphed on a second by second basis in real time). Through this physiological feedback, it is anticipated that the patient will learn how to control his/her bodily responses through mental means.

Biofeedback arose as an application of the learning theories of B.F. Skinner, Hull, Thorndike, Dollard and Miller, and John Watson. In particular, Neal Miller postulated that the established principles of learning that had so far been applied to overt behaviors could validly be applied to behaviors that were covert and presumed not under voluntary control.

Classical conditioning is also referred to as Pavlovian Conditioning after the seminal work of Russian scientists Pavlov and Sechenov in the early twentieth century. Classical conditioning is a laboratory learning paradigm by which a neutral stimulus (conditioned stimulus; CS) comes to elicit a new response (conditioned response; CR) by repeated pairing in close temporal proximity with another stimulus (unconditioned stimulus; UCS) that already elicits that response (unconditioned response; UCR). In subsequent presentations of the CS, the organism will then emit the UCR without pairing of the UCS. For example, the UCS might be food and the UCR is salivation, the CS, the ringing of a bell, is presented immediately prior in temporal pairing with the UCS, food. After repeated pairing of the ringing bell with food, the organism will come to salivate in response to the bell's ringing. The behaviors conditioned in this paradigm are typically unlearned, such as most physiological responses utilized in biofeedback practice. However, the learning paradigm most often appealed to as the theoretical underpinning of the field of biofeedback is not classical conditioning. Rather, biofeedback is generally considered a form of operant conditioning. This learning theory postulates that the consequence of a response changes the likelihood that the organism will produce that response again. The essential assumption of operant conditioning is that behavior is lawful and follows the rules of cause and effect and probability.

A basic supposition of operant conditioning is, if you wish a behavior to continue, you reinforce or reward that behavior. If you wish a behavior to decrease, or to stop completely, you do not reinforce that behavior. Thus, within the theoretical framework of operant conditioning, the main way that you strengthen a behavior is to follow it in close temporal proximity with a reward. The definition of a reinforcer is any stimulus change that occurs after a response and tends to increase the likelihood that a response will be repeated. It is important that the reinforcer follow the desired behavior quickly such that the delay in the presentation of the reward is kept to an optimally short delay. As the delay in the reward increases, the effectiveness of the reinforcer is generally decreased. There are many examples of positive reinforcement in our everyday life—receiving a bonus for outstanding work, receiving an A in a course for intensive studying, scoring a touchdown in a football game and the crowd's adulation.

In addition to positive reinforcement, there are three other possible consequences to behavior in operant

conditioning: (a) Negative reinforcement involves the removal of a consequence to a response that results in reduced likelihood that the behavior will be repeated in the future. (b) Positive punishment involves adding a consequence when a response is performed that serves to decrease the likelihood of the response occurring in the future. Examples are plentiful: child misbehaves, parent scolds child, child less likely to misbehave; drive over the speed limit, get ticket, less likely to speed; do poorly at work, get demoted, less likely to perform poorly on the job. Negative punishment involves the removal of a consequence to a behavior that serves to reduce the likelihood of that behavior occurring in the future. An examples is a parent playing with their child who is clearly enjoying the playtime; the child starts to yell loudly, the parent stops playing with the child, parent less likely to engage in positive play with the child. In clinical biofeedback applications, these other types of reinforcement contingencies are rarely used, with biofeedback clinicians preferring to use positive rewards to influence behavior.

Perhaps the most important principle in operant conditioning that directly involves clinical biofeedback training is that of shaping. Shaping is the learning process by which the predefined target response is achieved through gradual and systematic reinforcement. The training begins with a simple, existing response and basic criteria for reinforcement, with gradually more stringent criteria applied for reinforcement in order to achieve more complex and reliable responses. After the initial behavior is reliably performed, reinforcement is given contingent on the performance of more complex or difficult responses. This pattern of increasingly stringent contingent reinforcement continues until the final target behavior is achieved. Shaping can be assisted by modeling the skill to be learned before shifting the reinforcement schedule. In clinical biofeedback, the target behavior in shaping is often tailored to the individual learning style and abilities of the patient. For example, a patient undergoing EMG biofeedback training may be initially reinforced for detecting gross changes in electrical activity in a particular muscle group. Following the initial success, the reinforcement is tapered and made contingent on the patient being able to detect ever more subtle changes in muscular tension in the target region. This may continue until the patient is unable to further demonstrate more refined skill in detecting muscle tension.

There are other principles within operant conditioning that also apply to biofeedback practice. One of these involves discrimination training, in which the organism demonstrates the ability to differentiate between at least two stimulus conditions by emitting a different response to each stimuli. In clinical biofeedback, the concept of discrimination applies to the patients' ability to distinguish relatively subtle differences in physiological states. Another important principle, indeed the mortar that lays the foundation of clinical biofeedback training, is the concept of generalization: If a response is conditioned to one stimulus, the organism may also respond to a similar stimulus (generalization), but not to a dissimilar stimulus (discrimination). Discrimination learning is a goal in the early stages of biofeedback training, while generalization is a longer range goal. Clearly, the overarching aim of clinical biofeedback training is to take the learned process applied in the office setting and have that learning process come to apply to the everyday ''real world'' setting.

In addition to basic operant conditioning principles, there are general principles in basic psychophysiology that are important in clinical biofeedback training. The law of initial values states that the autonomic nervous system response to stimulation is a function of the prestimulus level (2). The higher the level of the response measure prior to a stressful stimulus being presented, the smaller the increase in response to the stressor, which is often referred to as a ceiling effect. Conversely, the higher the level of the measure prior to a relaxing stimulus being presented, the larger the decrease in response to the relaxing stimulus. When prestimulus response values are low prior to the presentation of a relaxing stimulus, this will lower the magnitude of the response and is often referred to as a floor effect. While the law has been shown to generally hold for measures of respiration and cardiovascular activity (such as heart rate and the vasomotor response), measures such as salivation and electrodermal response have not been found to be influenced by prestimulus values.

Homeostasis refers to the tendency of any organism to strive to maintain a state of equilibrium or rest. Homeostasis is believed to be maintained by a negative feedback loop, which is a theorized set of bodily mechanisms that provide information. This information directs the organism's physiological systems to decrease activity if levels of functioning are higher than normal, or to increase activity if levels are diminished relative to normal. Thus, all organisms strive to return to prestimulus levels of physiological arousal when presented with any stimulus.

Theories Underlying Clinical Biofeedback Training

There are two general theories underlying the use of biofeedback for most chronic benign medical disorders, such as anxiety, headache, musculoskeletal pain, and incontinence (3). The first is a direct psychophysiological theory, which attributes the etiology and/or maintenance of the disorder to specific physiological pathology. This biofeedback training modulates in a therapeutic direction. For example, it has traditionally been assumed that tension headache is caused by sustained contraction of skeletal muscles in the forehead, neck, and shoulder regions. Through the use of biofeedback, the patient learns to decrease muscle tension levels, leading to a decrease in headache activity. The second theory is predominantly psychological and postulates that there is a relationship between situational stress and the disorder in question. Through the use of biofeedback, the patient learns to regulate physiological responses such as muscle tension levels or sympathetic nervous system activity. This regulation leads to a decrease in overall stress levels, which brings about symptomatic relief. This amelioration of symptoms brought about by the learning of voluntary control of specific peripheral responses is postulated to be underpinned by central changes that occur along pain-relay and sympathetic pathways. For the case of headache treatment, this means that both muscle

relaxation and hand warming may indirectly dampen central brain mechanisms involved in the onset of headache. It is not necessary to view these theories as competing; they may be more appropriately viewed as complementary. Most clinicians subscribe to both theories, depending on the patient's presenting problem, clinical findings, and medical history.

Biofeedback Modalities and Instrumentation

Biofeedback instruments are first and foremost psychophysiological measuring instruments, to which has been added the capability to display the value of the measured parameter(s) in a form understandable to the subject. Feedback can be visual, auditory, or tactile.

The physiological measures generally employed in clinical biofeedback training are surface electromyographic activity (EMG) and skin surface temperature. Less often used physiological responses include measures of neuronal activity using electroencephalography (EEG), and measures of electrodermal response (skin resistance, skin conductance), cardiovascular activity (simple heart rate, heart rate variability, blood pressure, and vasomotor activity), and respiration (generally, respiration rate and depth).

Biofeedback instruments can be used to gain insight into a subject even if biofeedback therapy is not the goal. When assessment is the goal, care must be taken that the results are not contaminated by unintended biofeedback, such as the subject viewing the display. Subject spatial position with respect to windows, doors, and the professional must also be considered to avoid influencing the results. Environmental control of ambient temperature, humidity, and drafts is strongly recommended to minimize effects of these stressors on the subject.

Biofeedback instruments fall into three categories: research, clinical, and trainers. Research instruments are often configured from very flexible laboratory modules and their use is beyond the scope of this article. Many modern clinical instruments, however, are precise enough for basic or clinical research.

Clinical instruments are generally accurate, calibrated, and reliable. They are available as either stand-alone discrete units measuring a single parameter, or as computer-based multimodality systems. Some are comprehensive and accurate enough to be used for research. Trainers are single modality instruments intended to be purchased by or loaned/rented to the subject. They are less accurate and expensive than clinical instruments. Modern technology has made most trainers accurate and reliable despite their relatively low cost.

To protect both the subject and the professional, it is recommended that only FDA listed equipment be used. Even third-party software should meet this recommendation unless it is to be used only for educational purposes.

The biofeedback professional should have the proper academic credentials for the applications offered. Certification in use of the specific instrument modalities and applications is also recommended. The Biofeedback Certification Institute of America (BCIA), an affiliate of the Association for Applied Psychophysiology and Biofeedback (AAPB), offers certification in many areas. The American Physical Therapy Association (APTA) offers certification in EMG treatment of urinary incontinence. It is further recommended that the biofeedback professional receive training for the specific instruments used as there are some differences among instruments of the same type.

Most modern instruments are manufactured using standards traceable to the National Bureau of Standards (NBS). Modern electronic technology permits instruments to remain calibrated throughout their life, with the exception of sensors, which must be replaced occasionally. Instrument performance while attached to a subject, even the same subject at different times, can vary widely from improperly placed sensors, electromagnetic interference (EMI) and the subjects' condition. Even the professional is not a psychophysiological constant. Therefore it is strongly recommended that the using professional invest in test instruments and fixtures for each modality to be used so they can independently establish that the instrument(s) are operating correctly. The test device could be as simple as a laboratory thermometer used in a stirred water bath to verify temperature or a precision resistor to verify electrodermal accuracy. Biopotentials (EMG, EEG, ECG) require an electrode meter, and electrical safety, a volt Ω-ampere meter. Some commercial ECG signal simulators also provide sine and square-wave outputs, useful for testing EMG, EEG, and ECG instruments.

Circuitry and software of commercial instruments varies as to amplifier bandwidth, filter cutoffs, signal rectifiers, and integrators, which make it difficult to measure accurately direct comparisons between instruments of the same type from different manufacturers. Comparing results from different instruments of the same model from one manufacturer depends on the technology used. Older instruments with discrete component filters vary more than modern instruments due to component tolerances. Modern computer-based instruments employing software filters are more similar.

Professionals making comparisons with other professionals should use relative values such as percent decrease in finger temperature. A rough rule of thumb follows: Instruments vary about $\pm$ 10% across models and manufacturers, but subjects can vary as much as an order of magnitude in some measures. Lesson: Rely on known standard inputs to evaluate instrument performance.

Safety falls into three major areas: environmental, biological, and electrical. Environmental safety concerns trip hazards, sharp edges–corners, heaters, lamps and machinery in the subject–professional area. Biological safety concerns disinfecting re-useable sensors, subject chair, area, and so on. Disposable biopotential (EMG,EEG,ECG) electrodes are recommended. Until disposable EEG electrodes become available, they and all reusables should be cleaned and disinfected between uses. Electrical safety concerns both subject and professional. Battery powered instruments are intrinsically safe but if connected to a alternating current (ac), line operated device (i.e., computer, recorder, oscilloscope), through a nonisolated interface, they become a potential hazard.

Most computer-based instruments are isolated to stringent standards and provide complete electrical safety for

the connected subject. By isolating the computer system with a medical grade power transformer, the professional is also protected. Use the (recommended) ac current meter to verify isolation.

EMG Biofeedback Instrumentation

Biofeedback applications of EMG range from simple relaxation, using electrodes placed on the forehead, to complex neuromuscular retraining of stroke (cardiovascular accident, or CVA) victims utilizing four EMG channels on each of the affected and unaffected sides to retrain functional movements using both inhibition and reinforcement learning techniques.

The EMG signals for clinical biofeedback are the summation of muscle cell action potentials generated by the underlying muscles seen at the skin. They are acquired from surface electrodes applied over (or in the vicinity of) muscle(s) to be monitored. Signals from the electrodes are amplified and conditioned by high performance differential amplifiers. Signal characteristics of interest are in the range of 0.1–2000 μV amplitude, over a bandwidth of ~25–500 Hz. The amplified signal is then processed to a form suitable for display to the subject.

EMG Electrodes. Muscle signals are acquired using two (active) electrodes located along the muscle fiber axis. A third electrode (common, often erroneously called ground) establishes the instrument common at the subject common potential. Well-placed surface electrodes provide an adequate EMG signal to enable subjects to learn control and change in the desired direction (i.e., relaxation, EMG lower; reeducation, EMG higher).

Needle or wire subcutaneous electrodes can collect a more comprehensive representation of the EMG (motor and nerve cellular action potentials), but their use is limited to Neurology or research, due to the complexity and invasive nature of the procedures.

A surface electrode is a complex electrochemical network. The manufacturing process reduces and stabilizes all internal parameters, leaving only the electrolyte–subject skin interface for the clinician to cope with. Fortunately, EMG amplifier technology has considerably reduced requirements for electrode preparation (for most applications) to a good skin cleaning using alcohol or a commercial prep solution. A surface electrode basically consists of a contact resistance (i.e., 10–100 kΩ), paralleled by a capacitance (i.e., 1 nF), in series with a half-cell potential (i.e., 300 mV "battery").

The care required in electrode–lead–amplifier placement and mechanical stabilization depends on procedure dynamics and electrical environment considerations. For relatively static applications (i.e., relaxing in a chair), in a relatively benign electrical environment, much less preparation and virtually no stabilization are required. For dynamic (i.e., treadmill) applications in a hostile electrical environment (i.e., central urban), considerable care and expertise are required to obtain reliable signals. Proper electrode choice and stabilization means (i.e., taping) are required. Electrode mechanical stabilization is required in dynamic applications to minimize disturbing the half-cell potentials that form at the electrolyte–skin interface.

Other considerations for electrode selection and placement include surface curvature, movement, and sweating.

Today, most EMG biofeedback clinicians use disposable electrodes in consideration of disease and litigation problems. One newer type of surface electrode (hydrogel) can be moved between (properly prepared) sites on one patient during one session, but using any type of electrodes between subjects is not recommended.

EMG Amplifiers. The EMG amplifiers have benefited considerably from integrated circuit technology and today achieve performance inconceivable two decades ago. Many products still connect to electrodes with a cable or leads, but some amplifiers are small enough to mount directly on the electrode structure. While the amplifier characteristics are exceptional, the performance of the EMG channel can be compromised by asymmetries in electrodes and connecting leads, resulting in unequal electrical induction, causing artifact to appear as a differential error signal. The following are considered to be minimal specifications for a modern EMG amplifier:

- Low internal noise (<0.5 μV, p–p).
- High input impedance (Zin >100 MΩ).
- Flat bandwidth and sharp high and low frequency cutoffs (>18 dB/octave).
- High common mode rejection ratio (CMRR $> 10^7$).
- Common mode input range (CMR $> \pm 200$ mV).
- Static electricity shock protection (>2000 V).
- Gain stability (all causes) $> \pm 1\%$.

Bandwidths (3 dB) in common use for clinical biofeedback are 25–500 Hz, primarily used for neuromuscular reeducation training, and 100–200 Hz used for relaxation training and where bandwidth must be limited because of EMI considerations. A hardware or software switch provides bandwidth selection. The narrow bandwidth loses some of the EMG frequency spectrum, but is adequate for many applications, having the advantages of lower in-band noise and less artifact susceptibility. Some manufacturers believe that a single bandwidth of ~30–300 Hz captures most of the EMG signal and provides simplicity.

EMG Biofeedback Application Example. Consider the problem of acquiring a 1 μV signal from a muscle using two (active) surface electrodes and a reference (common) electrode. Characteristics of the reference electrode are not as critical as the actives because the high CMRR of the amplifier minimizes disturbances in its impedance and half-cell potential.

The two active signal acquiring electrodes are effectively back to back, in series with the (muscle) signal source impedance and generator. Impedance of the active electrodes is probably close, since they were applied in a like manner, and the very high input impedance of the amplifier makes differences negligible.

Active electrode half-cell potentials are a different matter. While they are similar, they are of such a great magnitude (i.e., 300 mV) compared to the signal (i.e., 1 μV) that small abrupt changes in either half-cell potential will be coupled through the amplifier as a large artifact, which is the reason why it is so important to mechanically stabilize surface electrodes. Small disturbances of electrodes and leads outside of the frequency band of interest are not seen.

Leads connecting electrodes to the amplifier must also be stabilized to reduce artifact. They should be twisted and taped down for dynamic applications. Triode (equilateral triangle group) electrodes allow placing the miniaturized amplifier directly on the electrode(s), reducing lead length to zero. Amplifier and lead mass must also be considered in dynamic applications.

In some applications, such as treadmill and (internal) pelvic floor applications, some artifact must be tolerated and either average or quiescent levels used for training.

Following successful acquisition and amplification of the EMG signal it is usually rectified in a precision operational circuit or subroutine to produce the time-variant average that is needed for quantification and display. Older instruments provide integral average and modern instruments provide root-mean-square (rms) average that is loosely held as an analog of power by some authors. A related instrument response parameter is the time constant (TC) of integration that follows rectification.

Neuromuscular reeducation applications require a short (i.e., 50 ms) TC to sense the slightest voluntary (phasic) response and relaxation training as long as one second, to provide a more slowly changing display promoting relaxation, when tonic levels are of more interest.

Therapists having detailed knowledge of the frequency spectral characteristics of the specific muscle(s) being trained view (on an oscilloscope) the raw (filtered but not rectified ac) signal for assessment and training effectiveness. Raw EMG is seldom used as feedback due to its' visual complexity.

Stand-alone instruments use either analogue operational or microprocessor implemented filtering, rectification, averaging, and display. Most modern instruments are microcomputer based. The biofeedback instrument manufacturer provides signal acquisition and conditioning hardware to be interfaced, and software to be installed in a PC. Most recent biofeedback software has been for use on Windows operating system PCs.

Some stand-alone instruments have data storage, statistical reporting, and downloading capabilities. Microcomputer-based systems provide very sophisticated data reporting capabilities, with raw data exporting for advanced analysis. Some systems feature general purpose software for more advanced users and application specific software for less sophisticated users performing repetitive procedures, such as pelvic floor muscle strengthening and synchrony training for urinary incontinence.

Temperature Biofeedback Instrumentation

Temperature biofeedback is primarily used to improve poor peripheral blood flow caused by chronic sympathetic arousal. Other medical conditions causing low peripheral blood flow must be ruled out prior to attempting biofeedback. Temperature is then taken as a partial integration of blood flow and thus of peripheral vasoconstriction caused by sympathetic arousal.

Training criteria are determined by the therapist for the particular subject, but digit temperature training goals range from 31.1 to 35 °C for most subjects.

There are several types of low cost "instruments" primarily used for group education. These include glass-alcohol thermometers, liquid-crystal (i.e., mood-dots and biotic bands), and digital thermometers. This section concerns itself with clinical grade instruments employing sensors, amplifiers, and displays.

Temperature Sensors. Small sensor(s) are affixed to finger(s) or toe(s) with porous surgical tape at nearly zero tension so as not to occlude blood flow, or provide a heat reservoir. The sensor(s) are in good thermal contact with, but electrically isolated from, the subject. Accuracy and linearity of modern commercial sensors are excellent, but it is essential to use sensors having a short thermal time-constant and minimal hysteresis so minute changes in temperature (i.e., blood flow) are immediately communicated to the subject.

Modern sensors are thermistors, compensated negative coefficient resistors that provide a linear negative change in resistance over the temperature range of 21–37.8 °C. A small direct current (dc) voltage is impressed across the thermistor resulting in increased current flow for increases in heat (due to increased blood flow) from the subject.

Temperature Amplifiers. Whether the temperature instrument is stand-alone or one modality of a multimodality system, the circuit consists of a compensated thermistor sensor, powered by a constant voltage, and a current-to-voltage amplifier that produces a temperature proportional voltage that is then displayed and/or digitized for further processing.

Amplifier characteristics depend on the thermistor resistance range, the excitation voltage, and the required output voltage. Bandwidths of 0–5 Hz provides adequate response and minimizes noise, improving resolution. Since signals are high level, a single-ended current amplifier can be used.

Absolute accuracy is important because both subjects and therapists tend to discuss readings in their respective groups for comparative purposes. Subjects particularly are very sensitive to numbers and an instrument differences could impede training progress.

Most instruments specify ± 0.56 °C absolute accuracy. It is often better than claimed and can be checked in a stirred waterbath against a standard lab mercury thermometer.

Resolution of most instruments is selectable for either 0.056 or 0.0056 °C resolution. Subjects tend to make large improvements in initial states of training and the coarser (0.056 °C) resolution is adequate. There is some question by these authors about whether the 0.0056 °C resolution has clinical utility.

Temperature Displays. A time-based line graph is the most common display, as it shows past history and current trend. Bar graphs, digital meters, and computer animation displays are also used. These should be selected by the therapist–subject team to be relevant to the subject's beliefs; and sometimes varied to keep motivation high.

Temperature Biofeedback Application. Sensor attachment is often the back of the finger or toe of interest using a breathable tape to prevent heat build-up. Tape tension should be just enough to affix the sensor without constricting blood flow. Lead length should be sufficient to allow the subject to achieve a comfortable position without lead tension.

Room temperature should be in the comfort range for the subject (i.e., 21.1–23.9 °C) and without any breeze. Relative humidity should be between 30 and 50% to avoid humidity stress or evaporation cooling.

A location free of transient noises (i.e., office activity, elevator, emergency vehicles) will facilitate subject focus and relaxation.

EEG Biofeedback Instrumentation

The EEG biofeedback instrumentation is similar to EMG biofeedback instrumentation in acquiring the biopotential signal. The reader is encouraged to review the section on EMG biofeedback instrumentation for a background on the following section.

Whereas EMG instrumentation acquires surface muscle signals at the skin, EEG instruments acquire signals on the scalp generated by brain cellular activity below. The EMG activity in this region is considered artifact and care must be taken in EEG signal processing to minimize EMG contamination. When large EMG artifacts cannot be removed from the EEG signal, feedback must be blocked or held constant until the EMG artifact has passed.

The EEG signal is acquired on the scalp using surface electrodes, generally different from EMG electrodes, in either a differential (bipolar) or referential (monopolar) mode. High performance differential amplifiers increase and condition the signals. Signals of interest are in the range of 0.5–100 V (p–p), over the frequency range of 1–30 Hz. Newer clinical research extends the frequency range to 1–50 Hz.

Modern EEG biofeedback instruments perform digital filtering on the amplified and conditioned EEG signal to obtain the frequency band(s) or full spectrum of interest (i.e., fast fourier transform, FFT).

Viewing the full frequency spectrum EEG signal is of interest in assessing brain state(s) and is sometimes used as feedback, displayed as bilateral (left and right hemisphere) displays back to back about as vertical centerline.

More often the feedback display is less complicated, consisting of vertical bars or lights representing several spectral bands logically combined to reward increases in the band of interest (i.e., beta, 16–20 Hz) and decreases in the band to be reduced (i.e., theta, 4–9 Hz). The EMG activity is also monitored and acts to inhibit or freeze the feedback if present.

This logic is also used to drive a computer generated graphic animation sequence (i.e., games) and/or audio feedback comprising several types of files and songs (i.e., wav, midi). Auditory feedback is useful for closed eyes training and can also be used to provide simultaneous reward-inhibit band information when the visual presentation does not provide it, as in the case of animation sequence visual feedback.

EEG Electrodes. The EEG electrodes are placed on the scalp in a standardized grid called the 10–20 system. Coordinates are determined with respect to the midline, between the nasion and the inion, and a line between the right and left ears, in percent of the distance from the reference axis.

Bipolar recordings are made between (2) active scalp electrodes, with a neutral site serving as amplifier common (i.e., center forehead). An active pair of electrodes is used for each additional site to be monitored. Only one common is required unless the amplifiers have isolated commons, in which case they must be tied to the subject, usually at the same neutral site.

Monopolar recordings are made with a single scalp electrode (+) with respect to a chosen reference site (−). Amplifier common is placed at a third (neutral) site. For bilateral monopolar recording, the left channel has a scalp electrode (+) somewhere on the left side of mid-line with its reference (−) on the left ear. Amplifier common could be either the left mastoid bone or forehead.

The right channel placement somewhat mirrors the left with the active scalp electrode (+) on the right side of midline, the active reference (−) on the right ear, and the common on the right mastoid bone or forehead. The left and right active electrodes are placed according to training objectives, not necessarily symmetrical.

Multichannel systems are monopolar having a scalp electrode for each of the (20) sites in the 10–20 system, all having the same (forehead) reference and instrument common.

An Electro-Cap having 20 scalp and one reference–common electrode, sized to the subject, is worn. Electrode sites are prepared through holes in each electrode. A cable harness connects the cap to a junction-box for connecting amplifiers to desired sites, or directly to the instrument, for the 20 channel system. Since all sites are measured referentially (i.e., monopolar), differential comparison between any sites can then be accomplished in software or circuitry.

It is important that both active electrodes (i.e., +, −) have the same electrode material and electrolyte, since the initial stage of the amplifier is dc coupled, and a large difference in half-cell potentials could saturate (block) the amplifier. The (2) active electrodes must be prepared similarly even if one is a scalp (cup or disk) placement and the other an ear-clip.

Gold, silver–silver chloride, and tin are common active electrode materials. The common electrode can be a different type, usually a disposable silver–silver chloride EMG electrode.

Electrode sites are prepared by mild abrading and infusing a conductive "prep" gel to stabilize and improve conductance of the scalp. This is followed by application of

an thixo-tropic electrolyte (i.e., 10–20 paste). The cup or disk electrode is then pressed on the paste until it is firmly sealed. Viscous force holds the scalp electrode in place during recording. Hair is managed with tape or cottonballs, which also helps retain and stabilize the placement. Ear clip electrodes are spring retained.

Reusable cup scalp and ear clip electrodes are in common use despite potential health and litigation problems. Several disposable systems have been or are being developed, but are not in wide use at this writing.

Careful electrode preparation will result in an impedance of <10 kΩ, usually measured using a 20 Hz ac impedance meter. The actual value required depends on the electrical environment of the treatment area. The EEG electrodes are generally unshielded, so electric induction is reduced with low impedance placements.

Electrode half-cell potentials must also be checked differentially with an electrode meter. Differences of more than ± 25 mV between the active electrodes usually indicates unstable placements or different materials used. If a different type of common electrode was used, a large half-cell potential difference between each active and the common is not problematical, as long as they are similar and stable.

EEG Amplifiers. Like EMG amplifiers, EEG amplifiers make good use of integrated circuit and surface-mount technologies to achieve performance never before seen. While amplifier characteristics are exceptional, the effective performance of the EEG channel can be compromised by asymmetries in electrodes, cables, and leads. The resulting imbalance between the plus (+) and minus (−) inputs to the differential amplifier can cause unequal electric induction producing an error that shows as increased noise or offset. Fortunately, most unwanted electric fields are outside the amplifier bandwidth. Still, good practice in electrode preparation, lead routing, and electrical environment control are necessary to avoid a setup that is electrostaticly "hot".

The following parameters are considered minimum for the modern EEG differential amplifier:

- Low internal voltage and current noise(<1 μV, 100 pA, p–p)
- High input impedance (Zin $>10^8$).
- Bandwidth (1–50 Hz).
- Frequency cutoffs (>18 dB/octave).
- High common mode rejection ratio ($>10^7$).
- Common mode input range (greater than ± 200 mV).
- Static electricity shock protection (>2000 V).
- Gain stability (all causes) greater than $\pm 1\%$.

The amplified EEG signal is then digitally filtered to produce the desired bands of frequencies to be used for assessment, training, or mapping.

EEG Displays and Feedback. The modern EEG instrument utilizes a computer to perform the signal processing, auditory and visual display generation, data collection–reduction, and reporting necessary for effective assessment and training. Today's sophisticated programs require a fairly high performance computer to perform all these tasks in apparent real time. Most manufacturers design around readily available Windows based personal computers having the following minimum characteristics:

Processor:	Pentium 3
Speed:	600 MHz
Memory	256 MB
Storage	10 GB
Display resolution	800×600
Video memory	32 MB ($4 \times$ AGP)
Operating system	Windows '98, or later
Auditory system	Sound Blaster Live
Speakers	Good quality with sub-woofer

Recent advances in clinical biofeedback have shown subjects can comprehend complex visual and multiparameter auditory feedback simultaneously. A higher performance computer with more working and video memory, and the very best auditory system is suggested for these applications.

A very useful feature of Windows'98 and later operating systems is the support of multiple display monitors. This makes it possible to stretch the display onto a second monitor. The therapist can construct displays having highly technical items on their monitor while the subject sees nontechnical items, such as animation.

A common use of the digitally filtered EEG signal is for treatment of Attention Deficit Disorder. Other applications using other bands of frequencies include hyperactivity and performance enhancement.

Full frequency spectrum displays are used to "quiet" a "noisy" brain under the guidance of a skilled therapist. Animation sequence displays accompanied by contingent auditory feedback provide effective training tools for the therapist.

Some 20 channel systems are capable of mapping the entire EEG frequency spectrum at every electrode location in the 10–20 coordinate system in (nearly) real time. The displays are updated in 50 ms providing a useful assessment mode.

EEG Biofeedback Application. The most important factor is the therapist's training. The EEG assessment and biofeedback treatment has progressed considerably beyond that required to use the less complex modalities of temperature, electrodermal, and EMG. Fortunately, few schools are keeping up with advances. The International Society for Neuronal Regulation (ISNR) is a good source to inquiry. The Neurofeedback division of the Association for Applied Psychophysiology and Biofeedback (AAPB) and ISNR have together addressed several critical issues in EEG biofeedback methodology and clinical application.

As with most psychophysiological assessment and biofeedback applications environmental temperature (21.1–23.9 °C), humidity (30–50%), illumination (as required), auditory noise (transient free, white noise is usually acceptable), and therapist's physical position with respect to the subject must be considered and controlled.

An electrically quiet environment is also desirable. Power lines, TV and radio antennas, and mobile communications are frequent sources of interference. Basement locations usually offer the electrically quietist locations and well-placed electrodes along with carefully routed leads provide the best immunity under the therapist's control.

Electrodermal Biofeedback Instrumentation

Electrodermal activity (EDA) is essentially a measure of palmar sweat gland activity of the fingers or hands. Bulk tissue impedance is largely ignored, being clinically less significant than the eccrine sweat gland activity, modulated by the sympathetic nervous system that modern instruments measure.

Electrodermal activity can be measured either as resistance, called skin resistance activity (SRA = SRL + SRR) or it's reciprocal skin conductance activity (SCA = SCL + SCR). The SCA has the advantage of being a linear function of the number of sweat glands conducting, while SRA has a hyperbolic relationship to the number of sweat glands conducting.

Electrodermal activity includes both the tonic (average) level (SRL or SCL) and short-term (phasic) response (SRR or SCR). Training goals for SCA are to lower the tonic level, indicative of chronic sympathetic arousal and reduce the phasic changes percent, indicative of over reactivity, either spontaneously or in response to stimuli.

The magnitude of phasic changes tend to occur as a percentage of the tonic level. For conductance, a 1 μS SCR change from a tonic level of 5 μS SCL is approximately equivalent to a 2 μS SCR change from a 10 μS tonic SCL level (i.e., 20%).

Most modern instruments measure conductance although one manufacturer recently reverted to resistance so that the same programmable amplifier could be used to measure any variable resistance sensor (i.e., temperature). Since conductance and resistance are reciprocals, either can be displayed with the conversion made in circuitry or software.

Subject phasic responses are delayed by 1–3 s neurophysiologically. This latency limits the usefulness of EDA as an early in-training feedback modality. It is an excellent assessment modality where the subject receives no feedback.

The EDA is a passive electrical parameter and must be elicited by impressing a small voltage on (conductance), or passing a small current through (resistance) the two electrodes, usually placed on the fingers. Either method is referred to as "excitation".

The normal range of EDA for human subjects is

Conductance:	0.5–50 μS
Resistance:	2 MΩ–20 kΩ

Most untrained subjects range between 2 and 10 μS. A value of 50 μS would be indicative of a soaking wet hand.

Electrodermal Electrodes. Normal placement of electrodes is on the palmar surface of the index and middle fingers of either hand, selection dictated by other modality sensors on a particular hand. Electrodes are held on the fingers by Velcro straps set for just enough tension to hold them on without causing blood "pulsing" or pounding in the fingers. Silver–silver chloride is the most common electrode material, but gold, stainless steel, and nickel-plated brass are also used.

Early instruments used monopolar (dc) excitation. At the values then used, electrode polarization and subsequent amplifier blocking tended to occur. Modern instruments use 10 μA current (resistance) or 200 mV (conductance) and reverse the excitation several times per second resulting in no net charge, thus avoiding electrode polarization. These values also limit current to 10 μA as required by the FDA.

Electrolyte between electrode and skin is not normally used clinically for EDA, but should be used, along with finger cleaning for clinical research or published studies.

Electrodermal Biofeedback Amplifier. Modern EDA amplifiers employ a switching technique to reverse electrode excitation several times per second to avoid electrode polarization. This requires a reversible voltage (conductance) or current (resistance) source, circuits easily implemented by operational amplifier techniques. Values are sampled after transient effects caused by the switching. The EDA is a relatively slowly changing modality and an effective amplifier bandwidth of 1–2 Hz is adequate.

The entire EDA (EDL + EDR) should be measured and later separated. Phasic changes (EDR) can be ac coupled to remove the EDL and amplified to produce the desired display sensitivity. This produces bipolar components to each phasic response (EDR), which can be difficult for subjects to understand.

Another method is to introduce an offset in the amplifier equal to the EDL resulting in only the phasic (EDR) component that can then be amplified separately. This method requires a relatively stable EDL or a circuit that samples the EDL and adjusts the offset continuously.

Electrodermal Feedback. Use of EDA as an assessment (without feedback) modality is highly recommended since it provides insight into the chronic sympathetic arousal level and reactivity of the subject.

Using EDA as a feedback modality in the beginning subject is discouraged as it is so nonspecifically reactive, subject relaxation and learning may be impeded. Advanced subjects, however, may find it challenging to control EDR while performing tasks.

For assessment, the time line graph of the entire EDA provides the most information on current and recent past history to the therapist.

The line graph, bar graph or contingent animation are all useful visual displays for the advanced subject.

Pitch-proportional (SCA) or pitch inversely proportional (SRA) auditory tones are useful for eyes-closed or task performance training. Care must be taken in choosing the pitch range to avoid alarming the subject with a "siren" effect.

Electrodermal Biofeedback Application. Silver–silver chloride electrodes are recommended. Low cost velcro-strap electrodes are affordable for each subject's course of training. Using these between subjects is not recommended from disease and therapist liability considerations. Disposable Ag/Ag CL EMG electrodes provide an alternative, but affixing them to digits is not as convenient.

Hand washing prior to the session is advisable for both uniformity and sanitation.

Use of electrolyte is not required for most clinical training, but should be used for clinical research or published studies. Saline gel or cream is suitable for most subjects. Non-saline gel is useful for allergic subjects. Hand washing following the session is good practice.

A small percentage (i.e., 7%) of subjects show little or no EDA. Always check the instrument with a known conductance–resistance standard (or the therapist) before concluding that the instrument is mal-functioning.

Since EDA is an elicited or exosomatic measure, care must be taken to place sensors from all used instrument modalities so that the current from EDA does not interfere with or impede other simultaneous measurements. Manufacture of multimodality biofeedback systems provide guidelines to minimize these situations. This is of particular concern if two EDA channels are used on the same subject.

As EDA is a relatively high level signal and excitation switching is slow, the modality is relatively impervious to electromagnetic interference.

Cardiopulmonary Biofeedback

Respiration (RSP) training has been shown to greatly improve the subjects' ability to relax and maintain self-regulation in the face of psychological and performance stressors. It tends to reduce performance anxiety and increases oxygen uptake and waste expulsion. It also increases peripheral circulation by reducing sympathetically activated vasoconstriction. Heart rate increase during inhalation is normal, implemented by the sympathetic nervous system. Para-sympathetic action slows heart rate during exhalation. Heart rate variability (HRV) training has also provided benefits in reducing rapid heart rate and tachycardia.

The HRV (measured as HR max - HR min) is frequently as high as 20 BPM in healthy 20 year-old adults, and decreases by age 50 often to ~10 BPM. Athletically active and physically well-conditioned individuals have higher variation in heart rate. Heart rate variability training aims at increasing the variability, and frequently increases it significantly higher than 20 BPM. Some authors have claimed that it goes as high as 50 BPM in peak training. Higher HRV is considered desirable in disease prevention and health promotion applications, and lower HRV correlates with cardiovascular morbidity and mortality. For example, lower HRV is a strong independent predictor of post-MI death (4).

By combining respiratory (RSP) and heart rate (HR) instrumentation, the interplay between respiration and heart rate, referred to as respiratory sinus arrhythmia (RSA), can be assessed and used as biofeedback to train subjects to optimize their natural cardiopulmonary rhythms under the influence of stressors.

Norms for HRV training have been published (5). One of the persistent problems in the field is the failure of researchers and practitioners alike to adhere to standard nomenclature and normative values for training. The following HRV frequency ranges have been established as standard for cardiopulmonary training:

Cardiac Rhythms. High frequency: 0.15–0.4 Hz; low frequency: 0.04–0.15 Hz; very low frequency: 0.0033–0.04 Hz; and ultra low frequency: <0.0033 (beyond clinical biofeedback measurement technology) (5).

Recently, a more careful examination of HRV has shown the very slow rhythms to be of interest in assessing dysregulation, other than that caused by the ANS, to be discussed in a separate HRV section, below. The very low frequency (VLF) range is to some extent correlated with dysregulation. Rhythms in the low frequency range are to some extent correlated with optimal homeostasis. Higher HR oscillations are found in rhythms in this range.

The RSP rates vary from 2 to 30 bpm requiring channel bandwidths of 0–5 Hz to faithfully reproduce the RSP waveform. Trained subject's RSP rates at rest are individually optimum and range from 6 to 9 pm.

The HR rates vary from ~40–180 bPM. Trained subject's HRs rest at ranges between 60 and 80 BPM The HRV (RSA) of 8–16 BPM as a function of RSP is normal and desirable.

Some instruments are capable of RSA assessment and biofeedback only. Instruments designed for HRV generally can also perform RSA procedures.

Respiratory Sinus Arrhythmia Instruments. Instruments for RSA have one or two respiration channels and one heart rate channel. Two respiration channels are desirable to train the subject to breathe abdominally, not thoracically.

The RSA instruments are computer implemented. Time line graphs of abdominal and thoracic RSP along with a beat-by-beat line graph of HR comprise an excellent assessment and training display. The line or (better) filled line graph is the most common display as it shows recent past history as well as current performance. Digital meters showing RSP and HR rates can be added, but the display tends to be too complicated, particularly for beginning subjects.

Raw data is saved, so the session can be replayed as desired. Statistics can be generated, but care must be taken to consider the effect of artifact. Both RSP and PPG HR channels are susceptible to movement artifacts. Some instrument software permits editing the raw data to minimize artifact contamination of statistics.

Respiration Sensors and Amplifiers. Respiration sensors for biofeedback comprise a stretchable segment and a belt or chain that is wrapped around the circumference of the abdominal and the thoracic regions of the subject. A slight prestretch (set at the point of maximum exhalation) allows the sensor to operate over the full range of circumference change caused by breathing. Care must be taken that restrictive clothing does not impede breathing.

Circumference change caused by breathing is a relative measure affected by subject breathing, posture, type of sensor, even temperature depending on the type of sensor used. Three types of sensors are in common use.

1. Rubber Bellows (Air Filled): Following placement the system is sealed, and sensor internal pressure changes with stretch. A transducer (half or full bridge strain gage) measures pressure changes. A dc coupled bridge amplifier, with offset control, provides amplification and the ability to "position" the output at the desired level for display and/or quantification. Typical pressure variations are on the order of ±15 mmHg (1.99 kPa) gage.
2. Tubular liquid-filled strain gage: An elastometric tube filled with a conductive thixo-tropic liquid. Changes in length and diameter of the tube, caused by breathing, varies the resistance of the gauge. Since the liquid is ionic, excitation (current or voltage) is reversed several times per second to prevent polarization. The amplifier comprises a reversible voltage or current source, gain, and offset capability to provide the desired output signal positively proportional to inhalation.
3. Magneto-position transducer: A magnetic armature is moved within an excited coil to produce a current proportional to movement. An elastomeric tubing provides a restorative force to track breathing movements. The amplifier converts the magnetically induced signal, provides gain and offset to provide a voltage positively proportional to inhalation.

Heart Rate Sensors and Amplifiers. Heart rate for clinical biofeedback is measured using either a finger photoplethysmograph (PPG) sensor or by acquiring the electrocardiogram (ECG) biopotential signal. The PPG sensor is easy to use, but its' output is subject to artifact from any disturbance of the placement. Acquiring the ECG requires placement of surface electrodes, but provides the virtually artifact-free signal that is required for reliable HRV analysis, and will be discussed in that section.

The PPG sensors employ an (invisible) infrared (IR) source of ~0.9 nm wavelength to illuminate the vascular bed of a finger (or toe). A photocell detects the backscatter caused by the opacity of blood at that wavelength. It is held in place with a Velcro strap or elastic band. The amplifier provides gain producing a pulse signal having information on HR, peripheral pulse amplitude (PPA), pulse volume (PPV) and rise and decay characteristics of the pulse.

Heart Rate Variability Instruments. Heart rate variability is useful for more than RSA training. Other factors, such as chemoreceptors, baroreceptors, the renin-angiotension system and various disease states affect HRV. Many of these variations occur at frequencies too slow to be perceived by or used as feedback for subjects.

Research has shown the frequency spectrum of HRV from 0.003 to 0.4 Hz to be useful in assessing disregulation of various systems. When HRV spectral components are distributed throughout the range of 0.003–0.4 Hz, HRV is said to be incoherent. In the coherent state, virtually all the energy occurs at one frequency in the range of 0.08–0.12 Hz and is individual specific.

The most comprehensive analysis of the slowest rhythms requires a 24 h data string of HR, such as is obtained in a Holter portable monitor of ECG, so various segments of Circadian rhythms can be analyzed. Data must be edited for artifact before spectral analysis by fast fourier transform (FFT) is performed. The comprehensive analysis can give insight into undiagnosed medical conditions.

In clinical psychophysiology, 5 min HRV data resolves frequencies from 0.003 to 0.4 Hz, adequate for observing ANS activity. Some instruments have added a 60 s HRV data gathering to provide 0.06–0.4 Hz spectral data, more easily obtained in the clinical situation, and adequate for determining coherence.

The ECG amplitude ranges between 0.3 and 2 mV for the QRS complex that is used to determine the interbeat interval from which the frequency spectrum is derived. The exact characteristics of the ECG signal are not as important in HRV applications as in clinical cardiology.

ECG Sensors and Amplifiers. The ECG is most reliably obtained by placement of chest electrodes using pregelled disposable Ag/AgCl sensors. For 24 h HRV studies, chest placement is mandatory.

In clinical psychophysiology, it is preferable not to remove clothing, so the wrist–wrist or wrist–ankle placement of sensors is preferred. Polarity of the ECG signal is important so the amplifier leads must be connected according to the manufacturer's instructions.

The EMG disposable sensors can be used to acquire the ECG in the clinic, but are not recommended for longer term studies as their adhesive surface area is considerably smaller than disposable ECG sensors, promoting half-cell disturbance artifact.

Typical ECG Amplifier Specifications:

Low internal noise (<2 μV p–p).
High Input Impedance [Zin> 10 MΩ].
Bandwidth (0.16–250 Hz).
Bandwidth cutoffs (>18 dB/octave).
Notch filter (60 Hz, in the United States).
Common mode rejection ratio [CMRR $> 10^7$].
Common mode input range (CMR ± 200 MV).
Static electricity shock protection (>2000 V).

Heart Rate Variability Instruments. The HRV instruments are computer implemented. They are also capable of performing RSA procedures using either ECG or PPG sensors to acquire HR.

Most HRV/RSA software is written for the Windows operating system. It is recommended that a fairly high performance computer be used to reduce HRV analysis computational time. A computer similar to that recommended for EEG is suitable.

Some instruments also support TMP and EDA in addition to the ECG, RSP, and PPG modalities.

One manufacturer offers a dual instrument interface making it possible for two computers to access one multimodality instrument, to perform simultaneous RSA/HRV, EEG, and other psychophysiological modality assessment and biofeedback procedures, with synchronized data collection. This instrument capability makes heart–mind interaction training and clinical research possible.

Heart Rate Variability Application. As with all biofeedback procedures establishing comfortable levels of temperature and humidity, with absence of transient auditory noise, is essential for focused, efficient, and reputable performance.

The Electromagnetic Interference (EMI) environment should meet the requirements of the most sensitive modality used (i.e., EEG).

Digitization of the ECG signal should be at least 256 $s \cdot s^{-1}$, with 512 $s \cdot s^{-1}$ recommended.

APPLIED CLINICAL EXAMPLES: TENSION AND MIGRAINE HEADACHE

We will now describe the typical EMG and hand surface temperature biofeedback procedures for tension and migraine headache, which we have used both clinically and in our research (3,6–8). As noted above, it is important to remember that when referring to EMG, the authors are alluding to surface electromyography, which uses noninvasive electrodes, is painless, and involves measuring the pattern of many motor action potentials; this is in contrast to EMG used as a diagnostic procedure in neurology, which uses invasive needle electrodes, measure the activity of a single motor unit, and is often quite painful.

The standard EMG biofeedback procedure for tension headache involves measuring the muscle tension in the frontalis muscle region by placing electrodes ~2.5 cm above each eyebrow and a ground electrode in the center of the forehead (9). The frontalis region has traditionally been assumed in clinical practice to be the best overall indicator of general muscular tension throughout the body. The standard thermal biofeedback training procedure involves attaching a sensitive temperature sensor, called a thermister, to a fingertip (usually the ventral surface of the index finger of the nondominant hand) with care taken not to create a tourniquet or inhibit circulation to this phalange.

EMG biofeedback is the modality most commonly used for tension headache, with the psychophysiological rationale being that muscle tension levels in the forehead, neck, and facial areas are directly causing or maintaining/exacerbating the headaches. It is also believed that individuals suffering from tension headache have high levels of stress and using EMG biofeedback as a general relaxation technique reduces their levels of stress, enabling tension headache sufferers to better cope with their headache activity.

Hand surface temperature biofeedback for migraine headache also has two possible mechanisms of action. The psychophysiological theory states that temperature biofeedback prevents the first of the two stages of migraine (vasoconstriction of the temporal artery and arterioles; the second stage is vasodilation, which causes the actual pain) from occurring by decreasing sympathetic arousal and increasing vasodilation to the temporal artery and arterioles. An alternative mechanism of action is the use of temperature as a general relaxation technique.

The EMG or temperature signal is then electronically processed using transducers to provide the patient with information on changes in the electrical activity of the muscles or surface skin temperature on a moment by moment basis. Generally, the signals are sampled every one-tenth of a second and integrated over the entire second. Both are quite sensitive, with the EMG sensor generally detecting changes of magnitude as low as a hundredth of a microvolt. The temperature sensor typically detects changes as low as one-tenth of a degree Fahrenheit. Through this feedback, the patient undergoing EMG biofeedback training learns how to relax the musculature of the face and scalp, and also learns how to detect early symptoms of increased muscle tension. In temperature biofeedback, the patient is taught how to detect minute changes in peripheral skin temperature, with the training goal being to increase hand temperature rapidly upon detection of low hand temperature. For EMG biofeedback, the feedback signal is usually auditory, and may consist of a tone that varies in pitch, a series of clicks that vary in frequency, and so on. Given the choice, >80% of patients receiving thermal biofeedback choose the visual display. The feedback display can be the pen on a voltmeter, a numeric output of the actual surface skin temperature, or a changing graph on a video screen. The format of the visual feedback display does not seem to affect learning or treatment outcome (10).

Common Type of Feedback Schedules in Clinical Applications. One of the challenges faced by both the biofeedback clinician and patient is selecting what type of feedback is most appropriate to facilitate learning to achieve rapid therapeutic benefit. There has been little research in this area; however, there are abundant anecdotal reports among the biofeedback community. There are three types of feedback schedules in clinical practice. By far, the most widely used method for delivering feedback is an analog display, which provides continuous information to the patient. For example, a tone that varies in relative pitch and frequency depending on an increase or decrease in the response being measured. However, in many applications, this may provide too much information to the patient, leading to information processing overload, retarding the learning process. A second type of feedback schedule employed in clinical practice is a binary display, where the patient receives information that is discrete, depending on achievement of a predefined training threshold. In threshold training, the feedback is turned on or off depending on whether the patient falls above or below the threshold. Threshold training is a clinical application of an operant shaping procedure, where the patient is reinforced for achieving successively closer approximations to the training goal. The third type of feedback schedule is an aggregate display of the training progress. In this type of feedback, the patient is given summary information at the conclusion of the treatment session (e.g., data averaged over each min interval in a 20 min training session). In

clinical practice, the integrated display of aggregate feedback is the most commonly applied training schedule.

Training to Criterion. Training to criterion is a term used by clinicians that involves continuing biofeedback training until the patient achieves a specified criterion of a learning end state. For example, biofeedback training will persist until the tension headache patient has demonstrated reduced muscle tension levels in the frontalis region to a stable 1-μV level. Although there is compelling logic behind this notion, there is little empirical data to support the practice of training to criterion. Exceptions to this are a report by Libo and Arnold (11) who found that every patient who achieved training criteria on both EMG and finger temperature also reported long-term improvement, and 73% of patients who did not improve failed to achieved training criterion in either modality. In another study, Blanchard et al. (12) presented data supporting the concept of training to criterion. They observed a discontinuity in outcome for migraine headache patients who achieve 35.6 °C or higher at any point during temperature biofeedback training. Those who reached this level had a significantly higher likelihood of experiencing a clinically meaningful reduction in headache activity (at least 50%) than those who reached lower maximum levels. This apparent threshold was replicated in a subsequent study (13). More representative of the research is a recent study by Tsai et al. (14), where they found no evidence to support the concept of training to criterion in a study of hypertensives. Fifty-four stage I or stage II hypertensives were taught thermal, EMG, and respiratory sinus arrhythmia biofeedback. Most participants (76%) achieved the thermal criterion; only 33 and 41% achieved the EMG and respiratory sinus arrhythmia criterion, respectively. Achievement of the criterion level in any of the three modalities was not associated with a higher improvement rate. These results contradict the notion that training to criterion is associated with clinical improvement.

Electrode Placement. An important consideration to be made by the clinician utilizing EMG biofeedback is at what sites to place the electrodes. This decision depends in large part on which of the two general theories underlying the use of biofeedback the clinician adheres to. In most instances, electrode placements appear not to matter. However, for tension headache this may not be the case.

Although the vast majority of published reports on tension headache utilize the frontalis region electrode placement, their is some controversy about this practice. This is perhaps because the Task Force Report of the Biofeedback Society of America, in their influential position paper on tension headache (15), strongly implied that frontal placement was the "gold standard" for biofeedback with tension headache sufferers, making no mention of other site placements. In the standard placement, muscle activity is detected not only in the forehead, but probably also from the rest of the fact, scalp, and neck, down to the clavicles (16).

Some writers (17,18) advocated attaching electrodes to other sites, such as the back of the neck or temporalis area, especially if the patient localizes his/her pain there. However, three of the four studies that compared biofeedback training from different sites between subjects found no advantage of one cite over the other (19–21). Arena et al. (22) published the only systematic comparison of a trapezius (neck and shoulder region) versus frontal EMG biofeedback training regimen with tension headache sufferers. They found clinically significant (50% or greater) decreases in overall headache activity in 50% of subjects in the frontal biofeedback group versus 100% in the trapezius biofeedback group. The trapezius biofeedback group was more effective in obtaining significant clinical improvement than the frontal biofeedback group. Thus, there is some limited support for the use of an upper trapezius electrode placement with tension headache sufferers. More research needs to be done in this area.

Discrimination Training. A concept in clinical biofeedback applications that is quite often discussed, particularly among those practitioners of EMG biofeedback training, is that of discrimination training. In this procedure, patients are taught to discriminate high levels of muscle tension from moderate and low levels. Feedback is given contingent upon successful differentiation among these varying levels of muscle tension. For example, a patient is asked to produce maximal muscle tension in a particular region, and given feedback reflective of this high level of muscle activity, followed by instruction to halve this level and consequent feedback. Then, finally, they are asked to halve this again, that is produce one-quarter of the initial level of muscle activity, followed by appropriate feedback reflecting success at this level. To our knowledge, there is little reliable data demonstrating that individuals specifically taught a muscle discrimination training procedure have clinical outcomes superior to those taught a standard tension-reduction method.

Sensitivity–Gain. The gain or sensitivity of the feedback signal is important to facilitate the training process in clinical biofeedback. Too high a gain may interfere with learning by providing indiscriminate feed back for extraneous responding, leading to frustration on the part of the learner. In addition, in many response measures, too high a sensitivity leads to increased artifact. Conversely, setting the gain too low leads to lack of feedback for responses that may be clinically meaningful, thereby interfering with the learning process. In clinical practice, there are established ranges in various applications, depending on the response measure employed, individual differences in patient responsivity, and the nature of the disease state. Sensitivity may be adjusted as needed using a shaping procedure. Some response measures involve more frequent changes in gain settings than others. For example, gain is frequently adjusted in EMG biofeedback applications, because the goal often is detection of quite subtle muscular activity changes, but infrequently changed in hand surface temperature training, where gross changes in skin temperature are usually necessary for clinical improvement.

Session Length and Outline. Treatment sessions usually last 30–50 min; 15–40 min of each session is devoted to the actual feedback training. In our research (and in our

clinical work), we have typically used the following format for biofeedback training sessions:

1. Attachment of electrodes and initial adaptation: 10 min.
2. In-session baseline, during which patients are asked to sit quietly with their eyes closed: 5 min.
3. Self-control 1, during which patients are asked to attempt to decrease their forehead muscle tension levels in the absence of the feedback signal: 3 min.
4. Feedback training, with the feedback signal available: 20 min.
5. Self-control 2, during which patients are asked to continue to decrease their forehead muscle tension levels in the absence of the feedback signal: 3 min.

The two self-control conditions are included to determine whether generalization of the biofeedback response has occurred. Generalization involves preparing the patient to, or determining whether or not the patient can, carry the learning that may have occurred during the biofeedback session into the "real world". If the patient can decrease muscle tension without any feedback prior to the biofeedback condition (Self-control 1 condition), then the clinician can assume that between-session generalization has occurred. If the patient can decrease their muscle tension without any feedback following the biofeedback condition (Self-control 2 condition), then the clinician can assume that within-session generalization has occurred.

There are other methods clinicians use to train for generalization of the biofeedback response. For example, in an attempt to make the office biofeedback training simulate real world situations, many clinicians initially train patients on a recliner; then, once they have mastered the rudiments of biofeedback in this extremely comfortable chair, they progress to, respectively, a less comfortable office chair (with arms), an uncomfortable office chair (without arms), and, lastly, the standing position. Finally, giving the patient homework assignments to practice the biofeedback response in the real world is a routine way of preparing them for generalization.

BRIEF REVIEW OF CLINICAL OUTCOME LITERATURE FOR BIOFEEDBACK

Anxiety Disorders–Stress Reduction

Biofeedback as a general relaxation technique has been in existence since the late 1960s. Indeed, it is common practice to call any form of biofeedback "biofeedback assisted relaxation", stressing the stress-reduction quality of the procedure. Where diagnoses are given, it is usually generalized anxiety disorder, although mostly the research defines anxiety or stress by global self-report measures or a simple paper and pencil instrument such as the Spielberger State-Trait Anxiety Inventory (i.e., scoring in the ninetieth percentile or above), rather than standard criteria such as the American Psychiatric Association's Diagnostic and Statistical Manual IV: revised (23). The primary modalities used for anxiety and stress reduction are EMG, hand surface temperature, and EEG. Nearly all the research has demonstrated that biofeedback is superior to placebo and wait-list controls for the treatment of stress and anxiety. There is some data to suggest (24) that EEG biofeedback to increase alpha waves may be superior to forehead EMG biofeedback and EEG biofeedback to decrease alpha waves in terms of decreasing heart rate activity, but not in terms of decreasing self-reported anxiety levels, where their were no differences between the three groups. When biofeedback has been compared to relaxation therapy, there is no difference between the two treatments in terms of their clinical efficacy (25).

One typical study was that of Spence (26). He took 55 anxious subjects, and gave them either electrodermal response, hand surface temperature, or forehead EMG biofeedback based on a pretreatment psychophysiological assessment (subjects were given feedback corresponding to that physiological parameter that changed the most during stress). All groups reported significant reductions in their anxiety symptoms, and 15 months later 76% of subjects were still symptom-free for anxiety, regardless of the type of feedback they received.

Moore (27) reviewed the EEG biofeedback treatment of anxiety disorders and pointed out that there are many limitations in the research to date. Unfortunately, many of his concerns hold for the EMG and temperature biofeedback literature as well, such as comparisons to relevant placebos, examination of such factors as duration of treatment, type and severity of anxiety, and so on.

Tension and Vascular (Migraine And Combined Migraine–Tension) Headache

By a large margin, the greatest number of articles supporting the efficacy of biofeedback for any disorder in the clinical treatment literature pertains to its use with vascular and tension headache. For both types of headache, biofeedback has been shown to be superior to both pharmacological and psychological placebo, as well as wait list control, in numerous controlled outcome studies. Biofeedback for headache is usually compared to relaxation therapy or cognitive therapy (a form of psychotherapy focusing on changing an individual's pain-and stress-related self-statements and behaviors). Arena and Blanchard have recently reviewed the biofeedback treatment outcome literature on tension and vascular headache (3,7,8).

With tension headache, the biofeedback approach used is EMG (muscle tension) feedback from the forehead, neck, and/or shoulders. For relaxation therapy alone, successful tension headache treatment outcomes generally range from 40 to 55%, for EMG biofeedback alone, this value ranges from 50 to 60%, and for cognitive therapy, from 60 to 80%; when EMG biofeedback and relaxation are combined, the average number of treatment successes improves from $\sim$50 to $\sim$75%; when relaxation and cognitive therapy are combined, success increases from 40 to 65%. When compared to relaxation therapy, there is usually comparable efficacy.

For patients with pure migraine headache, hand surface temperature (or thermal) is the biofeedback modality of choice, and it leads to clinically significant improvement in

40–60% of patients. Cognitive therapy by itself gets ~50% success. A systematic course of relaxation training seems to help when added to thermal biofeedback (increasing success from ~40 to 55%), but cognitive therapy added to the thermal biofeedback and relaxation does not improve outcome on a group basis. Relaxation training alone achieves success in from 30 to 50% of patients, and adding thermal biofeedback boosts that success (from ~30 to 55%). There appears to be a trend in the literature for thermal biofeedback to be superior to relaxation therapy.

For patients with both kinds of primary benign headache disorders (migraine and tension type), the results with thermal biofeedback alone are a bit lower, averaging 30–45% success; relaxation training alone leads to 20–25% success. Thermal biofeedback consistently appears to be superior to relaxation therapy with combined headache. The best results come when thermal biofeedback and relaxation training are combined. With this combination treatment, results show 50–55% success rates (adding thermal biofeedback to relaxation raises success from 20 to 55%; adding relaxation therapy to thermal biofeedback increases success from 25 to 55%). Most experts strongly recommend a combination of the two treatments for these headache sufferers.

Lower Back Pain

Arena and Blanchard (7) recently reviewed the biofeedback literature for low back pain and concluded that biofeedback appears to hold promise as a clinically useful technique in the treatment of patients with back pain. While the evidence indicates that optimal clinical improvement is clearly obtained when biofeedback is used within the context of a comprehensive, multidisciplinary pain management program, the cumulative weight of the evidence suggest that EMG biofeedback is likely to be helpful, as a single therapy, in the treatment of musculoskeletal low back pain, obtaining success rates of from 35 to 68% improvement on follow-up.

However, there were many concerns about the literature. Only two studies have directly compared biofeedback to relaxation therapy, and both of these studies were significantly flawed so as to limit definitive conclusions. Direct comparisons of biofeedback to relaxation therapy are clearly needed. Longer (at least 1 year) and larger scale (at least 50/group) follow-up studies are required. Evaluations of treatments based on diagnosis (i.e., the cause of the pain) should be conducted. Comparisons of various biofeedback treatment procedures, such as paraspinal versus frontal electrode placement, or training while supine versus training while standing, are necessary. Finally, further evaluations of patient characteristics predictive of outcome, such as gender, race, chronicity, psychopathology, and psychophysiological reactivity, are needed.

Myofascial Pain Dysfunction

Myofascial pain dysfunction (MPD) syndrome, also known as temporomandibular joint (TMJ) syndrome, is considered a subtype of craniomandibular dysfunction that is caused by hyperactivity of the masticatory muscles. It is characterized by diffuse pain in the muscles of mastication, mastication muscle tenderness, and joint sounds and limitations. Although disagreement exists as to the cause of the hyperactivity (e.g., occlusal problems vs. psychological stress), several researchers have examined the use of EMG biofeedback as a treatment, which can provide relief by teaching patients to relax the muscles of the jaw. Consistent with the logic of this approach, the most common electrode placement is on the masseter muscle, although frontal muscle placements have also been used. Excellent overviews of the treatment of MPD syndrome can be found (28,29).

Arena and Blanchard (7) recently reviewed the MPD biofeedback literature and noted that, although the majority of the studies had significant limitations, when taken as a whole they appeared to be quite impressive in support of the efficacy of EMG biofeedback for MPD syndrome. EMG biofeedback is at least as effective as traditional dental treatments such as occlusal splint therapy. Curiously, it was noted that no MPD syndrome study of biofeedback as a treatment in and of itself had been published since 1989. Given the extremely positive results, this observation is somewhat perplexing.

Deficiencies in the research on biofeedback treatment for MPD syndrome are similar to those discussed in the lower back pain section, above. Large scale outcome studies are needed, comparing (a) masseter versus frontal versus temporalis placement sites; (b) biofeedback versus relaxation; (c) biofeedback versus traditional dental strategies, and, (d) biofeedback in conjunction with other treatments versus traditional dental strategies. The latter approach has been used by Turk and co-workers (30–32), in a number of recent, methodologically elegant studies. In these studies various combinations of biofeedback, stress management training, and intraoral appliances were used, with results showing strong support for combined treatments. Finally, lack of long-term follow-ups, or for that matter, any follow-up at all, is a serious limitation that needs to be corrected.

Fibromyalgia

There have been a number of studies examining the efficacy of EMG biofeedback in the treatment of fibromyalgia (see Arena and Blanchard (5), for a review of the studies before 2000). The majority of the studies concluded that EMG biofeedback is useful in reducing fibromyalgic pain. Fibromyalgia is a type of nonarticular, noninflammatory rheumatism that is characterized by diffuse pain, sleep disturbance, tenderness, and functional impairment. Three studies have been published since 2000. Mueller et al. (33) gave 38 fibromyalgia patients EEG biofeedback, noted statistically significant decreases in pain, mental clarity, mood, and sleep. Van Santen et al. (34) compared physical fitness traing to EMG biofeedback and usual treatment on 143 female patients with fibromyalgia. They found no difference between the three groups on any measure. Recently, Drexler et al. (35) broke 24 female fibromyalgia patients down into those with abnormal psychological test (MMPI) results and those with normal psychological test profiles. Psychologically abnormal

individuals were helped more by the biofeedback training than were psychologically normal individuals. Given the relatively promising results [all five of the pre-2000 studies (36–40) obtained positive results], it appears that large scale, controlled EMG biofeedback studies looking at factors such as psychological profiles and gender would now be appropriate.

Biofeedback for Gastrointestinal Disorders: Constipation Pain, Irritable Bowel Syndrome, Urinary, and Fecal Incontinence

The biofeedback literature on treatment of constipation pain, especially in children, is both impressive and growing. In adults, Jorge et al. (41) recently reviewed the literature and noted that, overall, mean percentage of success is 68.5% for studies that examine constipation attributable to paradoxical puborectalis syndrome. Mason et al. (42) examined 31 consecutive patients who received biofeedback training for idiopathic constipation. Twenty-two of the patients felt subjectively symptomatically improved. They noted that the symptomatic improvement produced by biofeedback in constipated patients was associated with improved psychological state and quality of life factors.

In the constipation pain literature regarding children, three studies particularly stand out. Benninga et al. (43) gave 29 children who suffered from constipation and encopresis an average of five sessions of EMG biofeedback of the external anal sphincter. At 6 weeks, 55% were symptom free. Another group of investigators (44) placed 13 children who suffered from constipation into a standard medical care group, while another group of 13 children were placed in a EMG biofeedback (of the external anal sphincter–from 1 to 6 sessions) plus standard medical care group. At 16 month follow-up, all children were significantly improved, with the biofeedback plus standard medical care group significantly more improved than the standard medical care only group.

One large scale study, however, does not support the efficacy of EMG biofeedback for constipation pain. In a procedure similar to Cox et al. (44), van der Plas et al. (45) placed 94 children who suffered from constipation into a standard medical care group, while another group of 98 children were placed in a five-session EMG biofeedback (of the external anal sphincter) plus standard medical care group. At 18 month follow-up, over one-half of the children in both groups were significantly improved, with no significant difference between the two groups. In spite of this large scale study suggesting no advantage to the inclusion of EMG biofeedback to conventional medical care, we believe that there is sufficient evidence to conclude that EMG biofeedback is a useful technique in treating the pain of both adult and childhood constipation, especially when the patient has proven refractory to standard medical care.

Biofeedback for irritable bowel syndrome has been in existence since 1972, but nearly all of the studies are small and uncontrolled. The type of feedback is generally thermal biofeedback, however, two groups have used novel feedback approaches with some success. Leahy et al. (46) have developed an electrodermal response biofeedback device that uses a computer biofeedback game based on animated gut imagery. This significantly reduced symptoms in 50% of 40 irritable bowel syndrome patients. Radnitz and Blanchard (47), using an electronic stethoscope placed on subjects' abdomens, gave bowel sound biofeedback to five individuals with irritable bowel syndrome. Three of the five patients had reductions in their chronic diarrhea by over 50% (54, 94, and 100%). Results were maintained at 1- and 2-year follow-up (48). Large scale controlled outcome studies comparing biofeedback to pharmacological and dietary interventions for irritable bowel syndrome symptoms need to be conducted.

Biofeedback For Cancer Chemotherapy Effects

Biofeedback has been used to decrease the negative side effects of cancer chemotherapy, especially the anticipatory nausea. While biofeedback assisted relaxation does seem to help these patients, biofeedback by itself (i.e., not using a relaxation emphasis), while reducing physiological arousal, does not reduce the anticipatory nausea. This is an area where relaxation therapy seems to have a clear advantage over biofeedback. For example, Burish and Jenkins (49) randomly assigned 81 cancer chemotherapy patients to one of six groups in a 3 (EMG biofeedback/skin temperature biofeedback/no biofeedback) $\times$ 2 (relaxation/no relaxation) factorial design. They concluded, "The findings suggest that relaxation training can be effective in reducing the adverse consequences of chemotherapy and that the positive effects found for biofeedback in prior research were due to the relaxation training that was given with the biofeedback, not the biofeedback alone" (p. 17).

Biofeedback for Cardiovascular Reactivity: Hypertension, Raynaud's Disease, and Cardiac Arrhythmia

Biofeedback has been used as a treatment for essential hypertension since the late 1960s. The type of feedback used is either direct blood pressure feedback or temperature biofeedback. There appears to be no difference in terms of clinical outcomes between the two biofeedback modalities. In a recent influential meta-analysis of 22 randomized controlled outcome studies, Nakao et al. (50) found that biofeedback resulted in averaged blood pressure decreases of 7.3/5.8 mmHg (0.97/0.77 kPa) compared to clinical visits or nonintervention controls. It resulted in averaged blood pressure decreases of 4.9/3.5 mmHg (0.65/0.46 kPa) compared to sham or nonspecific behavioral interventions. Statistical analysis indicated that, after controlling for the effects of initial blood pressures, biofeedback decreased blood pressure more than nonintervention controls, but not more than sham or nonspecific behavioral interventions. Further analyses revealed that when the treatments were broken down into two types, biofeedback assisted relaxation, as opposed to simple biofeedback that did not offer other relaxation procedures, was superior to sham or nonspecific behavioral intervention. Nakao et al. (50) concluded that, "Further studies will be needed to determine whether biofeedback itself has an antihypertensive effect beyond the general relaxation response" (p. 37).

It has long been believed that temperature biofeedback is more efficacious than medication in the treatment of Raynaud's disease (51,52). Raynaud's disease is a disease of the peripheral circulatory system that is caused by insufficient blood supply to the hands and feet. It can result in cyanosis, numbness, pain, and, in extreme cases, gangrene and subsequent amputation of the affected finger or toe. The vasospastic attacks are triggered by cold and, to a lesser extent, anxiety and stress. Recent data, however, has failed to transparently support the belief that temperature biofeedback is a more effective treatment than medication for Raynaud's disease.

The Raynaud's treatment study (53) was a large, multicenter randomized controlled trial comparing sustained relief nifedipine, pill placebo, temperature biofeedback, and EMG biofeedback (a behavioral control) on 313 individuals diagnosed with primary Raynaud's disease. Results indicated that while nifedipine was significantly different from medication placebo in reducing vasospastic attacks, temperature biofeedback was not significantly different from psychological placebo (EMG biofeedback) in reducing vasospastic attacks. Comparison of nifedipine and temperature biofeedback indicated a nonsignificant ($p = 0.08$) trend for the nifedipine to result in greater reductions in vasospastic attacks. However, 15% of the nifedipine group had to discontinue the treatment due to adverse reactions to the medication. The interpretation of the biofeedback results of the Raynaud's treatment study, however, have been criticized by the behavioral investigators of the project (54). They note that a substantial proportion of subjects in the temperature group (65%) did not achieve learning, compared to only 33% in a normal comparison group who achieved successful learning by the end of the 10 biofeedback sessions in the protocol.

EEG Biofeedback (Neurofeedback)

Ramirez et al. (55) exhaustively reviewed the scientific literature on EEG biofeedback treatment of Attention Deficit Disorder (ADD). These authors conclude that, as in many other areas of clinical biofeedback practice, the positive evidence from anecdotal sources and case reports is plentiful, but a dearth of rigorous studies does not allow firm inferences to be drawn about the therapeutic efficacy of enhanced alpha wave activity and hemispheric lateralized biofeedback training. The EEG biofeedback training with a combined training goal of modifying the pattern of theta and beta wave activity has shown promising implications for management of ADD in adults. Studies using the theta/beta training paradigm have reported significant improvement in academic, intellectual, and adaptive behavioral functioning following EEG treatment. Other studies using sensorimotor rhythm training (recording from the "Rolandic" cortex) have produced behavioral and cognitive improvements in ADD patients. Unfortunately, these studies like those in other therapeutic areas of biofeedback are plagued with methodological problems including small sample sizes, absent or inadequate placebo controls, no randomization to treatment conditions, and insufficient follow-up of patient status. However, some authors of recent nonrandomized studies contend that EEG biofeedback shows promising evidence of therapeutic efficacy on the core symptoms of childhood ADHD in comparison to or in combination with standard stimulant medication therapy, family counseling, and an individualized educational intervention (56).

Another clinical problem in which EEG biofeedback was tested in the early 1970s was for control of frequent and disabling seizures. These studies have been reviewed by Lubar (57). The most common types of EEG recording and feedback training successfully studied in human subjects are EEG alpha rhythm (8–13 Hz) recorded from the occipital region of the brain, theta activity (4–7 Hz), and beta activity (>14 Hz). With the introduction over the recent decades of effective and relatively safe antiepileptic drugs, interest in systematic research and clinical application of EEG biofeedback as a nonmedication method of seizure control has waned. However, intractable seizures are still encountered in routine clinical practice despite all pharmacotherapeutic efforts. Implantable stimulatory devices and surgical interventions are reserved for highly selected patients and carry significant risks. For those patients with uncontrolled seizure disorder who have been unresponsive to standard anticonvulsant medication regimens and/or are not candidates for surgical treatment, Lubar has advocated that they be considered for a trial of sensorimotor rhythm EEG biofeedback training for the most common types of psychomotor seizures (57). Note that the equipment is expensive and the training procedures are complex and time consuming, and thus practitioners familiar with EEG biofeedback treatment of epilepsy may be difficult to find.

Quantitative EEG recording and specialized biofeedback training protocols have been developed and tested in the treatment of addictive disorders such as alcoholism. Peniston et al. (58) studied a protocol for enhancing EEG alpha and theta wave activity, and improving "synchrony" among the brain wave rhythms along the power spectrum. Peniston et al. (58) propose that alcoholics have a predisposition to "brain wave desynchrony" and deficient alpha activity and show a vulnerability to alcohol's capacity to produce reinforcing (pleasant and relaxing) levels of slow brain wave activity. These investigators have evaluated the treatment in a series of studies suggesting that their neurotherapy protocol reduces subjective craving among severe alcohol abusers, improves psychological functioning on personality measures, increases alpha and theta activity levels, increases beta-endorphin levels, and increases time to relapse. However, a large, independent randomized controlled trial of Neurotherapy did not show the incremental benefit to relapse prevention of adding electronic neurotherapy to a traditional residential treatment program for severe, chronic alcoholics (59). Although widely practiced, the clinical utility and theoretical rationale of EEG biofeedback in treating alcohol abusers remains controversial among the scientific biofeedback community. While promising data exists to suggest the potential role of EEG biofeedback in substance abuse treatment, further research is needed to illuminate the conceptual basis of such treatment and the reliability of clinical improvements for alcoholism and other addictive disorders.

There are very limited data from controlled studies on the use of EEG biofeedback for control of symptoms associated with Tourette syndrome, a behavioral impulse control disorder characterized by a constellation of motor and vocal tics (involuntary behaviors). A few scattered case reports describe positive results using a course of EEG sensorimotor rhythm biofeedback training to treat complex motor tics and associated attention deficit symptoms (60). There may be overlap in this treatment approach with the observation that epilepsy cases with motor involvement show some remediation following sensorimotor EEG biofeedback training. There is speculation and anectodal reports to suggest that anxiety, depression, and attentional symptoms associated with complex tics in Tourette's may be the most responsive targets for psychophysiological treatment. Because of the multiple symptom clusters of Tourette's, and the likelihood that different treatment protocols are needed to address the range of affected behaviors, focusing treatment on the whole condition is difficult, and often a prioritization of the most severe problems must occur to serve as the focus of clinical attention. Most patients are managed on a medication regimen by their physicians and EEG biofeedback is seen a useful adjunct in selected cases that have not responded adequately to pharmacologic management alone or where medication usage is to be reduced for some reason.

Cardiovascular Reactivity

Heart rate variability (HRV) biofeedback is being studied as a psychophysiological means of managing heart problems such as cardiac arrhythmia. Earlier isolated attempts at biofeedback interventions with cardiovascular ailments using simpler unitary heart rate (beats/min) or blood pressure (mmHg) measures have not been remarkably successful on modifying disease states. Heart rate variability is derived from the standard deviation of the beat-to-beat time interval (in ms) recorded in the laboratory with an ECG machine or with a Holter monitor using 24 h ambulatory monitoring methods. Heart rate variability has been proposed as a more robust metric of overall cardiac health in that it provides an indirect marker of the heart's ability to respond to normal regulatory impulses that affect its rhythm. With higher levels of HRV, it is proposed that there is a better balance between the combined sympathetic and parasympathetic inputs to the heart. Generally, greater HRV is associated with relaxed states and slow or regular breathing pattern. The HRV biofeedback training is claimed to offer a more precise method for helping clients to moderate the heightened sympathetic activity that is associated with elevated stress, anxiety, and dysphoric mood. Relatively greater levels of heart rate variability have been associated with better heart health. Biofeedback of breathing rate and depth is also used to increase respiratory sinus arrhythmia, which may be connected to therapeutic increases in HRV. A few small-scale studies have been conducted that show the clinical potential of HRV biofeedback in cardiovascular diseases (61); however, the results are inconsistent, and methodological problems abound with the existing studies. As yet, there is little evidence from larger scale randomized controlled trials conducted at independent laboratories demonstrating the therapeutic efficacy of HRV biofeedback in specific cardiovascular disease states such as cardiac arrhythmia.

Incontinence: Urinary and Fecal

EMG biofeedback training of the bladder-urinary sphincter and pelvic floor musculature has been found to be an efficacious intervention for urge urinary incontinence, especially among female geriatric populations, and is usually related to destrusor muscle contraction instability or reduced bladder volume (62). Some form of Kegel exercises are often used to train the muscles of the pelvic floor that are in continuity with the external urethral sphincter. Biofeedback with behavior modification training of the anorectal-anal sphincter musculature along the pelvic floor has been reported to be successful in treatment of fecal incontinence of various etiologies (63). Small, insertable EMG sensors are usually used in current treatment protocols for urinary incontinence in female patients and for fecal incontinence. A second EMG channel with abdominal placement is often recommended to better isolate contraction of the pelvic muscles from activity of accessory muscle of the legs, buttocks, and abdomen during the training exercises. Some degree of voluntary contractibility of sphincter muscles and rectal sensitivity are necessary for successful biofeedback treatment. While biofeedback training for urinary incontinence has a longer history of usage, and thus a larger empirical base (64), there is considerable evidence to suggest the efficacy of EMG biofeedback in a majority of adult patients with fecal incontinence (65). Unfortunately, there was a great deal of variability in biofeedback instrumentation used among these studies, treatment protocols followed, and outcome measures with uncertain validity.

Stroke and Mild Traumatic Brain Injury

There is very limited research in the area of EEG biofeedback in treatment and rehabilitation of the neurological impairments resulting from stroke or closed head injury. There are a number of anecdotal reports and small case series that suggest a place for quantitative EEG analysis in the functional assessment of neurological symptoms secondary to stroke and head injury. A highly individualized QEEG protocol used in these studies is sometimes called EEG entrainment feedback recording from the surface of brain regions suspected to be pathologically involved in the functional impairments (66,67). Neuromuscular reeducation is a general term used to describe assessment and treatment methods that may include EMG biofeedback and are applied to helping neurologically impaired patients (such as poststroke patients) with regaining gross motor functions necessary for carrying out activities of daily living and ambulation. In a meta-analysis of controlled trials using EMG biofeedback for neuromuscular reeducation in hemiplegic stroke patients, the authors concluded that EMG biofeedback resulted

in significant functional gains (68). While these results are promising, the specific effects of EMG biofeedback in stroke rehabilitation remain unclear as some of the studies reviewed included other interventions such as physical therapy or gait training as part of the rehabilitation program.

Sexual Dysfunction

Surface EMG biofeedback of targeted abnormalities in pelvic floor musculature are implicated in the pathogenesis of vulvovaginal pain (vulvodynia) syndromes such as dsypareunia resulting from vulvar vestibulitis. These have been successfully used in the stabilization of pelvic floor muscles leading to 83% reduction of pain symptoms, improved sexual function, and psychological adjustment at 6-month follow-up (69). As in other EMG biofeedback protocols for assessment and modification of abnormal pelvic floor musculature, an individualized assessment is performed to identify the patient's specific neuromuscular abnormality, with subsequent biofeedback training designed to modify the muscle tension and contractile weakness of the target muscles. However, gynecological surgery (vestibulectomy), on average, appears to produce superior outcomes (70).

FUTURE DIRECTIONS OF BIOFEEDBACK THERAPY

There are five areas that biofeedback research and clinical work are heading or should focus on. They are (*1*) expanding upon and refining current research; (*2*) applications of biofeedback and psychophysiological assessment to the "real world" environment (i.e., ambulatory monitoring); (*3*) applications of biofeedback training to new populations; (*4*) applications of biofeedback to applicatuons of biofeedback to the prinany the primary care setting; (*5*) alternative methods of treatment delivery.

1. Expanding Upon and Refining Current Research. Although biofeedback is considered a mature treatment, there are surprisingly many basic areas that need to be explored further. Such basic questions as (a) whether massed versus distributed practice produces greater physiological learning, (b) whether the presence or absence of the therapist in the room retards or enhances the acquisition of the biofeedback response, (c) the usefulness of coaching, (d) is their any value in training to criterion, (e) whether group biofeedback enhances or retards psychophysiological learning or clinical affects clinical outcome, have not been satisfactorily answered. Moreover, with the notable exception of headache, nearly every area is missing large scale treatment outcome studies (i.e., 25 or greater subjects per condition), in which biofeedback treatment is compared to placebo, another psychological treatment, conventional medical treatment, and so on. Many studies fail to describe the instrumentation and biofeedback procedures sufficiently to allow replication of the research. Often diagnostic criteria are not given, or diagnoses are commingled (e.g., conduct disorder children with attention deficit disorder children, or generalized anxiety disorder with simple phobias). Such failures to answer basic research questions or to conduct research in a scientifically acceptable manner are troubling and need to be corrected.
2. Applications of Biofeedback and Psychophysiological Assessment To the "Real World" Environment (i.e., ambulatory monitoring). Biofeedback clinicians have attempted to use their psychophysiological monitoring equipment to assist in setting treatment goals and to further explore the relationship between the mechanism of action believed to be involved in the underlying pathology of the disorder in question. For example, many clinicians use ambulatory blood pressure monitors to determine when their hypertensive patients are most reactive (work, driving, etc.) and tailor exercises to be maximized around those situations of elevated blood pressures. Use of such ambulatory equipment for other responses such as EMG, hand temperature, and respiration would be quite useful and such studies need to be performed.

   There have been only a few studies examining the relationship between ambulatory monitoring in the naturalistic environment and the presumed pathological response underlying the disease, with the exception of bruxism and temporomandibular joint dysfunction, where measurement in the natural environment by telemetry, portable tape recording, and digital EMG integration have been reported (71). Unfortunately, with those exceptions, when such studies have been conducted, they have arrived at negative findings, quite possibly due to the difficulty in reducing the data and inability to control all the relevant variables (sleep is a relatively controlled environment). For example, Hatch et al. (72) had 12 tension headache subjects and nine nonheadache controls wear a computer-controlled EMG activity recorder in their natural environment for 48–96 consecutive hours. The EMG activity of the posterior neck or frontal muscles was recorded 24 h/day. During waking hours, subjects rated their perceived levels of stress, pain, and negative affect at 30-min intervals. The EMG activity of headache and control subjects did not differ significantly, and EMG activity did not covary with stress, pain, or negative affect. Cross-correlations among EMG activity, pain, and stress revealed little evidence of leading, contemporaneous, or lagging relationships. Interrupted time series analysis showed no consistent muscle hyperactivity during a headache attack compared to a headache-free baseline period.

   Arena and co-workers designed a portable activity monitor for simultaneously recording and quantifying surface EMG signals and body movements in the natural environment (73). Two independent channels record EMG activity from symmetric muscle groups to determine contraction patterns. The EMG signals are amplified, filtered, integrated for 1 s, and converted to a digital value. Full scale was

jumper selected to accommodate a wide range of muscle activity. Electrode resistance >20 kΩ generates an alarm to signal poor contact or lead-off condition. The EMG voltages less than a preset threshold are not integrated.

The movement sensors are electrolytic potentiometers whose output are proportional to angular position and linear acceleration. The outputs are differentiated and summed to obtain angular acceleration with minimal response to linear movement. The peak value and 1 s integral are converted to digital values.

Subjective evaluation of pain and activity may be annotated by a 16-button keypad. An hourly auditory alarm reminds the user to enter subjective evaluations.

Data is saved in static random-access memory in binary coded 3-byte words. Power is supplied by a 9 V alkaline battery and converted to ±5 V by switching regulators. At the end of 18 h of recording, all power is turned off except for standby power to memory. A low-battery condition will also switch power to the standby mode. Data retained in the standby mode is valid for at least 7 days.

Arena et al. demonstrated that the device is highly reliable in 26 healthy controls (74). They then had 18 tension-type headache subjects and 26 control subjects wear the device attached to the bilateral upper trapezius muscles for 5 consecutive days for up to 18 h a day (75). During waking hours, subjects rated their perceived levels of stress, pain, and physical activity at 60-min intervals. Similar to Hatch, the EMG activity of headache and control subjects did not differ significantly, and EMG activity did not covary with stress, pain, or physical activity levels. Examination of cross-correlations among EMG activity, pain, physical activity, and stress revealed little evidence of an isomorphic, precursor or consequence relationship. There were no consistent differences between a headache and nonheadache state on muscle activity levels.

Arena et al. (75) concluded that there were so many variables entering into the natural environment, that use of such devices required a sophistication not available to the average clinician or researcher, and that treating headache patients nomothetically as an aggregate, rather than idiographically, as individuals, may also present difficulties. For example, some individuals may lie down as soon as a headache begins, while others may continue with their daily routine. Other individuals may have consequence, precursor, or isomorphic relationships between their head pain, except the changes occur on an every 5 min basis rather than a 1 h basis. With still others, to identify a relationship 5 consecutive days is not enough. Given the fact that the technology has increased exponentially to allow much more sophisticated data reduction and statistical analysis, we feel that the time is now ripe for a renaissance in such an area of research, which has been dormant for nearly a decade.

3. Applications of Biofeedback Training to Other Populations. As biofeedback is considered to be an established field, investigators have begun to take the treatments and expand it to other, similar clinical problems. For example, in the field of headache, biofeedback has been shown to be effective with the elderly, children, and pregnant women (6). Areas that need to be explored further to determine whether biofeedback treatment effects can be generalized are headaches in depressed individuals, headaches in individuals following cerebral aneurysm, headaches due to eyestrain, posttraumatic headache, and headache in multiple sclerosis patients. Similarly, the anxiety disorders literature needs to be expanded to include children, the elderly, anxiety due to a medical condition, and so on.
4. Application in Primary Care Medical Settings. Because of the growing recognition of the high prevalence of psychosomatic and psychophysiological disorders that present in primary care settings, the increased availability and implementation of psychophysiological assessment and biofeedback interventions in these healthcare settings appears to be timely (76). Many behavioral medicine interventions including biofeedback may be more efficiently and effectively delivered in these primary practice settings as the focus of these interventions is often toward the goals of preventing or slowing disease progression rather than treating severe or complicated problems that are well established. This approach is in contrast to conventional practice where patients with complicated medical problems or cooccurring psychological symptoms are referred out to specialty behavioral medicine clinics or other specialists (e.g., physical therapists) for psychophysiological treatment. As many chronic health problems are progressive in nature, by the time referral is made, the patient's condition is likely to have worsened to the point where behavioral intervention including biofeedback training may have far less impact than had it been instituted earlier in the disease course.

However, for biofeedback to be successfully integrated into the busy primary care office practice setting, certain modifications will have to be made in the context of assessment protocol and treatment delivery. First, behavioral assessment will have to be brief, but informative and practical, yielding results that are helpful to the primary care team in managing the patient's medical problems. The assessment results will have to be readily incorporated into the medical record of the patient rather than assigned privileged status, as mental health records frequently are, and therefore accessible to few if any providers for reference in primary care delivery. Second, the psychophysiology assessment and biofeedback treatment program will have to be carefully standardized and mid-level providers such as nurses, physician assistants, psychology technicians, or other mental health therapists trained in the competent and efficient delivery of these services. A doctoral level psychologist on staff or

consulting from another facility should be available to supervise these services to monitor quality and assess outcomes. Third, the rapid advancement of biofeedback equipment in terms of measurement accuracy, increased reliability with precision electronics, lowered cost, much improved portability through miniaturization, and enhanced patient convenience with alternative sensor technology has enabled the possibility of almost entirely home-based, self-administered treatment. Instruction and support can be provided by nursing staff, consultant psychologist, or other health professional through less frequent office visits and telephone consultation as needed. Arena and Blanchard have recently outlined in greater detail steps one should take to apply behavioral medicine techniques such as biofeedback to the trement of headaches in the primary care setting (3).

5. Alternative Methods of Treatment Delivery. The availability of relatively low cost, high precision biofeedback training devices lends itself to the possibility of a limited-therapist-contact, largely home-based treatment regimen. Blanchard and co-workers published three separate studies (77–79) evaluating a treatment regimen of three sessions (>2 months) combining thermal biofeedback and progressive relaxation training. In all three instances, very positive results were found for this attenuated form of treatment. Similar results were reported by Tobin et al. (80).

   We believe that some limited therapist contact is often necessary, so that patients understand the rationale for the treatment and that problems (trying too hard, thermistor misplacement, etc.) can be caught and corrected early. We also believe that detailed manuals to guide the home training, and telephone consultation to troubleshoot problems, are crucial in this approach. Given the national push for improving the efficiency of treatments, this approach has much to recommend it. We should also note that this home-based approach was not as successful as office-based treatment of essential hypertension with thermal biofeedback (81).

   This limited therapist contact does not have to be face-to-face with the therapist, however. It can be conducted via the Internet or using a videoconferencing telemedicine application. Devineni and Blanchard (under review 82) conducted a randomized controlled study of an Internet-delivered behavioral regimen composed of progressive relaxation, limited biofeedback with autogenic training, and stress management in comparison to a symptom monitoring waitlist control. Thirty-nine percent of treated individuals showed clinically significant improvement on self-report measures of headache symptoms at post-treatment. At 2-month follow-up, 47% of participants maintained improvement. There was a 35% within-group reduction of medication usage among the treated subjects. The Internet program was noticeably more time cost-efficient than traditional clinical treatment. Treatment and follow-up dropout rates, 38.1 and 64.8%, respectively, were typical of behavioral self-help studies.

   Arena et al. (83) recently reported a small ($n = 4$) uncontrolled study investigating the feasibility of an Internet and/or telemedicine delivery modality for relaxation therapy and thermal biofeedback for vascular headache. Each subject was over the age of 50 and had suffered from headaches for >20 years. Subjects came into the clinic for treatment but never saw the therapist in person. Instead, all treatment was conducted through the use of computer terminals and monitors. The only difference between this treatment and office-based treatment was the physical presence of the therapist. Results indicated one of the subjects was a treatment success (>50% headache improvement, and two others had between 25 and 50% improvements. Thus, it seems that further exploration into the potential of telemedicine and internet delivery of psychophysiological interventions is warranted.

In summary, an attempt has been made to review the basic theories underlying the application of biofeedback training to the amelioration of a broad range of general medical and psychiatric disorders. Also covered are the main types of biofeedback systems, technical specifications of instrumentation, and engineering design considerations for major functional components of biofeedback apparatus. A sampling of the many clinical problem areas in which psychophysiological assessment technology and biofeedback instrumentation have been utilized with varying degrees of success is discussed. This coverage of instrumentation is not exhaustive. Given the rich basic science underpinnings of biofeedback and the wide appeal among both health professionals and the general public, this coverage was by necessity selective and opportunistic. Throughout this article are found references to key resources of primary literature and authoritative reviews of the biofeedback field. This technological area is one of the most promising of both professional psychology practice and consumer oriented general healthcare. The field of biofeedback is far from matured, with medical application of basic research finding beginning only ~30 years ago. The field is probably in its second generation with a more rigorous examination of its scientific underpinnings, casting aside unproven or implausible theories related to disease process and treatment efficacy, and development of a sound empirical basis for its assessment methods and interventions. We are witnessing the continued rapid advancement of microcomputer technology and digital electronics coupled with the accumulation of knowledge of the basic mechanisms involved in human health and disease. There is a great potential for an evolution of biofeedback from its early origins with focus on nonpathological states and development of body awareness, human potential, and wellness to a more refined and sophisticated understanding and application of techniques toward the maintenance of health and prevention of disease states that is seamlessly integrated into the individual's lifestyle.

BIBLIOGRAPHY

Cited References

1. Olton DS, Noonberg AR. Biofeedback: Clinical applications in behavioral medicine. Englewood Cliffs, NJ.: Prentice Hall; 1980.
2. Wilder J. The law of initial values. Psychosom Med 1950;12: 392–400.
3. Arena JG, Blanchard EB. Assessment and treatment of chronic benign headache in the primary care setting. In: O'Donohue W, Cummings N, Henderson D, Byrd M, editors. Behavioral integrative care: Treatments that work in the primary care setting. New York: Allyn & Bacon; 2005. p 293–313.
4. Carney RM, Blumenthal JA, Stein PK, Watkins L, Catellier D, Berkman LF, Czajkowski SM, O'Connor C, Stone PH, Freedland KE. Depression, heart rate variability, and acute myocardial infarction. Circulation 2001; 104:2024–2028.
5. Task Force of the European Society of Cardiology and The North American Society of Pacing and Electrophysiology. Heart rate variability: Standards of measurement, physiological interpretation, and clinical use. Eur Heart J 1996;17:354–381.
6. Arena JG, Blanchard EB. Biofeedback training for chronic pain disorders: A primer. In: Gatchel RJ, Turk DC, editors. Chronic pain: Psychological perspectives on treatment. 2nd ed. New York: Guilford Publications; 2002. p 159–186.
7. Arena JG, Blanchard EB. Biofeedback Therapy for Chronic Pain Disorders. In: Loeser JD, Turk D, Chapman RC, Butler S, editors. Bonica's Management of Pain. 3rd ed. Baltimore: Williams & Wilkins; 2001. p 1755–1763.
8. Blanchard EB, Arena JG. Biofeedback, relaxation training and other psychological treatments for chronic benign headache. In: Diamond ML, Solomon GD, editors. Diamond's and Dalessio's The Practicing Physician's Approach to Headache. 6th ed. New York: W. B. Saunders; 1999. p 209–224.
9. Lippold DCJ. Electromyography. In: Venables PH, Martin I, editors. Manual of Psychophysiological Methods. New York: John Wiley & Sons; 1967.
10. Evans DD. A comparison of two computerized thermal biofeedback displays in migraine headache patients and controls. [Unpublished dissertation]. State University of New York at Albany; 1988.
11. Libo LM, Arnold GE. Does training to criterion influence improvement? A follow-up study of EMG and thermal biofeedback. J Behav Med 1983;6:397–404.
12. Blanchard EB, Andrasik F, Neff DF, Saunders NL, Arena JG, Pallmeyer TP, Teders SJ, Jurish SE, Rodichok LD. Four process studies in the behavioral treatment of chronic headache. Behav Res Ther 1983;21:209–220.
13. Morrill B, Blanchard EB. Two studies of the potential mechanisms of action in the thermal biofeedback treatment of vascular headache. Headache 1989;29:169–176.
14. Tsai P, Calderon KS, Yucha CB, Tian L. Biofeedback training to criteria and blood pressure reduction. Proceedings of the 34th Annual Meeting of the Association for Applied Psychophysiology and Biofeedback. Wheat Ridge, Colorado: AAPB; 2003.
15. Budzynski T. Biofeedback in the treatment of muscle-contraction (tension) headache. Biofeedback Self Regul 1978; 3:409–434.
16. Basmajian JV. Facts vs. myths in EMG biofeedback. Biofeedback Self Regul 1976;1:369–378.
17. Belar CD. A comment on Silver and Blanchard's (1978) review of the treatment of tension headaches by EMG biofeedback and relaxation training. J Behav Med 1979;2:215–218.
18. Hudzinski LG. Neck musculature and EMG biofeedback in treatment of muscle contraction headache. Headache 1983;23:86–90.
19. Hart JD, Cirhanski KA. A comparison of frontal EMG biofeedback and neck EMG biofeedback in the treatment of muscle-contraction headache. Biofeedback Self Regul 1981;6:63–74.
20. Philips C. The modification of tension headache pain using EMG biofeedback. Behav Res Ther 1977;15:119–129.
21. Philips C, Hunter M. The treatment of tension headache. II. Muscular abnormality and biofeedback. Behav Res Ther 1981;19:859–489.
22. Arena JG, Bruno GM, Hannah SL, Meador KJ. A comparison of frontal electromyographic biofeedback training, trapezius electromyographic biofeedback training and progressive muscle relaxation therapy in the treatment of tension headache. Headache 1995;35:411–419.
23. American Psychiatric Association. Diagnostic and Statistical Manual of Mental Disorders IV-TR. Washington, DC: APA Press; 2000.
24. Rice KM, Blanchard EB, Purcell M. Biofedback treatments of generalized anxiety disorder: preliminary results. Biofeedback Self Regul 1993;18:93–105.
25. Eppley KR, Abrams AJ, Shear J. Differential effects of relaxation techniques on trait anxiety: a meta analysis. J Clin Psychol 1989;45:957–974.
26. Spence J. Maximization of biofeedback following cognitive stress pre-selection in generalized anxiety. Percept Mot Skills 1986;63:239–242.
27. Moore NC. A review of EEG biofeedback treatment of anxiety disorders. Clin Electroencephalog 2000;31:1–6.
28. Crider AB, Glaros AG. A meta-analysis of EMG biofeedback treatment of temporomandibular disorders. J Orofac Pain 1999;13:29–37.
29. Gevirtz RN, Glaros AG, Hooper D, Schwartz MS. Temporomandibular disorders. In: Schwartz MS, editor. Biofeedback: A practitioner's guide. 2nd ed. New York: Guilford; 1995. p 411–428.
30. Turk DC, Rudy TE, Kubinski JA, Zaki HS, Greco CM. Dysfunctional patients with temporomandibular disorders: Evaluating the efficacy of a tailored treatment protocol. J Consult Clin Psychol 1996;64:139–146.
31. Turk DC, Zaki HS, Rudy TE. Effects of intraoral appliance and biofeedback/stress management alone and in combination in treating pain and depression in patients with temporomandibular disorders. J Prosthet Dent 1993;70:158–164.
32. Greco CM, Rudy TE, Turk DC, Herlich A, Zaki HH. Traumatic onset of temporomandibular disorders: Positive effects of a standardized conservative treatment program. Clin J Pain 1997;13:337–347.
33. Mueller HH, Donaldson CC, Nelson DV, Layman M. Treatment of fibromyalgia incorporating EEG-Driven stimulation: a clinical outcomes study. J Clin Psychol 2001;57: 933–952.
34. van Santen M, Bolwijn P, Verstappen F, Bakker C, Hidding A, Houben H, van der Heijde D, Landewe R, van der Linden S. A randomized clinical trial comparing fitness and biofeedback training versus basic treatment in patients with fibromyalgia. J Rheumatol 2002;29:575–581.
35. Drexler AR, Mur EJ, Gunther VC. Efficacy of an EMG-biofeedback therapy in fibromyalgia patients. A comparative study of patients with and without abnormality in (MMPI) psychological scales. Clin Exp Rheumatol 2002; 20:677–682.
36. Nolli M et al. Evaluation of chronic fibromyalgic pain before and after EMG-BFB training. Algos 1986;3:249–253.

37. Ferraccioli G et al. EMG-biofeedback training in fibromyalgia syndrome. J Rheumatol 1987;14:820–825.
38. Waylonis GW, Perkins RH. Post-traumatic fibromyalgia: A long-term follow-up. Am J Phys Med Rehabil 1994;73:403–412.
39. Buckelew SP et al. Self-efficacy predicting outcome among fibromyalgia patients. Arthritis Care Res 1996;9:97–104.
40. Sarnoch H, Adler F, Scholz OB. Relevance of muscular sensitivity, muscular activity, and cognitive variables for pain reduction associated with EMG biofeedback for fibromyalgia. Percept Mot Skills 1997;84:1043–1050.
41. Jorge JM, Habr-Gama A, Wexner SD. Biofeedback therapy in the colon and rectal practice. Appl Psychophysiol Biofeedback 2003;28:47–61.
42. Mason HJ, Serrano-Ikkos E, Kamm MA. Psychological state and quality of life in patients having behavioral treatment (biofeedback) for intractable constipation. Am J Gastroenterol 2002;97:3154–3159.
43. Benninga MA, Buller HA, Taminiau JA. Biofeedback training in chronic constipation. Arch Dis Child 1993;68:126–129.
44. Cox DJ, Sutphen J, Borowitz S, Dickens MN, Singles, Whitehead WE. Simple electromyographic biofeedback treatment for chronic pediatric constipation/encopresis: Preliminary report. Biofeedback Self Regul 1994;19:41–50.
45. Van der Plas RN et al. Biofeedback training in treatment of childhood constipation: A randomized controlled trial. Lancet North Am Ed 1996;348:776–780.
46. Leahy A, Clayman C, Mason I, Lloyd G, Epstein O. Computerized biofeedback games: a new method for teaching stress management and its use in irritable bowel syndrome. J R Coll Phys London 1998;32:552–556.
47. Radnitz CL, Blanchard EB. Bowel sound biofeedback as a treatment for irritable bowel syndrome. Biofeedback Self Regul 1988;13:169–179.
48. Radnitz CL, Blanchard EB. A 1- and 2-year follow-up study of bowel sound biofeedback as a treatment for irritable bowel syndrome. Biofeedback Self Regul 1989;14:333–338.
49. Burish TG, Jenkins RA. Effectiveness of biofeedback and relaxation training in reducing side effects of cancer chemotherapy. Health Psychol 1992;11:17–23.
50. Nakao M, Yano E, Nomura S, Kuboki T. Blood pressure-lowering effects of biofeedback treatment in hypertension: A meta-analysis of randomized controlled trials. Hypertens Res 2003;26:37–46.
51. Freedman RR. Long-term effectiveness of behavioral treatments for Raynaud's Disease. Behav Ther 1987;18:387–399.
52. Sedlacek K, Taub E. Biofeedback treatment of Raynaud's Disease. Prof Psychol Res Pr 1996;27:548–553.
53. Raynaud's Treatment Study Investigators. Comparison of sustained-release nifedipine and temperature biofeedback for treatment of primary Raynaud phenomenon. Results from a randomized clinical trial with 1-year follow-up. Arch Intern Med 2000;24:1101–1108.
54. Middaugh SJ, Haythornthwaite JA, Thompson B, Hill R, Brown KM, Freedman RR, Attanasio V, Jacob RG, Scheier M, Smith EA. The Raynaud's Treatment Study: biofeedback protocols and acquisition of temperature biofeedback skills. Appl Psychophysiol Biofeedback 2001;26:251–278.
55. Ramirez PM, DeSantis D, Opler LA. EEG biofeedback treatment of ADD: A viable alternative to traditional medical intervention? Adult Attention Deficit Disorder: Brain Mechanisms and Life Outcomes. In: Wasserstein J, et al., editors. Ann. N. Y. Acad. Sci. New York: New York Academy of Sciences; 2001.
56. Monastra VJ, Monastra DM, George S. The effects of stimulant therapy, EEG biofeedback, and parenting style on the primary symptoms of attention-deficit/hyperactivity disorder. Appl Psychophysiol Biofeedback 2002;27:231–249.
57. Luber JF. Electroencephalographic biofeedback methodology and the management of epilepsy. Integr Physiol Behav Sci 1998;33:1053–1088.
58. Peniston EG, Kulkosky PJ. Neurofeedback in the treatment of addictive disorders. In: Evans JR, Abarbanel A, editors. Introduction to Quantitative EEG and Neurofeedback. San Diego: Academic Press; 1999. p 157–179.
59. Taub E, Steiner SS, Weingarten E, Walton KG. Effectiveness of broad spectrum approaches to relapse prevention in severe alcoholism: A long-term, randomized, controlled trial of Transcendental Meditation, EMG biofeedback, and electronic neurotherapy. Alcohol Treat Q 1994;11:187–220.
60. Tansey MA. A simple and a complex tic (Gilles de la Tourette's syndrome): Their response to EEG sensorimotor rhythm biofeedback training. Int J Psychophysiol 1986;4:91–97.
61. Brody C, Davison ET, Brody J. Self-regulation of a complex ventricular arrhythmia. Psychosom: J Consult-Liaison Psychiat 1985;26:754–756.
62. Burgio KL, Locher JL, Goode PS, Hardin JM, McDowell BJ, Dombrowski M, Candib D. Behavioral vs. drug treatment for urge urinary incontinence: A randomized controlled trial. J Am Med Assoc 1998;280:1995–2000.
63. Jorge JMN, Habr-Gama A, Wexner SD. Biofeedback therapy in the colon and rectal practice. Appl Psychophysiol Biofeedback 2003;28:47–61.
64. Tries J, Brubaker L. Application of biofeedback in the treatment of urinary incontinence. Prof Psychol Res Pr 1996;27:554–560.
65. Norton C, Kamm MA. Anal sphincter biofeedback and pelvic floor exercises for faecal incontinence in adults—A systematic review. Aliment Pharmacol Ther 2001;15:1147–1154.
66. Rozelle GR, Budzynski TH. Neurotherapy for stroke rehabilitation: A single case study. Biofeedback Self Regul 1995;20:211–228.
67. Byers AP. Neurofeedback therapy for a mild head injury. J Neurother 1995;1:22–37.
68. Schleenbaker RE, Mainous AG. Electromyographic biofeedback for neuromuscular reeducation in the hemiplegic stroke patient—A meta-analysis. Arch Phys Med Rehabil 1993;74:1301–1304.
69. Glazer HI, Rodke G, Swencionis C, Hertz R, Young AW. Treatment of Vulvar Vestibulitis Syndrome with Electromyographic Biofeedback of Pelvic Floor Musculature. J Reprod Med 1995;40:283–290.
70. Bergeron S, Binik YM, Khalife S, Pagidas K, Glazer HI, Meana M, Amsel R. A randomized comparison of group cognitive-behavioral therapy, surface electromyographic biofeedback, and vestibulectomy in the treatment of dyspareunia resulting from vulvar vestibulitis. Pain 2001;91:297–306.
71. Burger C, Rough J. An EMG integrator for muscle activity studies in ambulatory subjects. IEEE Trans Biomed Eng 1983; 66–69.
72. Hatch JP, Prihoda TJ, Moore PJ, Cyr-Provost M, Borcherding S, Boutros NN, Seleshi E. A naturalistic study of the relationships among electromyographic activity, psychological stress, and pain in ambulatory tension-type headache patients and headache-free controls. Psychosom Med 1991;53:576–584.
73. Searle JR, Arena JG, Sherman RA. A portable activity monitor for musculoskeletal pain disorders. Proc Annu Int Conf IEEE Eng Med Biol Soc. 1989.

74. Arena JG, Bruno GM, Brucks AG, Searle JD, Sherman RA, Meador KJ. Reliability of an ambulatory electromyographic activity device for musculoskeletal pain disorders. Int J Psychophysiol 1994;17:153–157.
75. Arena JG, Bruno GM, Brucks AG, Searle JD, Sherman RA, Meador KJ (unpublished manuscript). The measurement of surface EMG in tension-headache subjects in the natural environment: Ambulatory recordings of data from five consecutive days.
76. Gatchel RJ, Oordt MS. Future trends and opportunities. In: Gatchel RJ, Oordt MS, editors. Clinical Health Psychology and Primary Care: Practical Advice and Clinical Guidance for Successful Collaboration. Washington, DC: American Psychological Association; 2003.
77. Blanchard EB, Andrasik F, Appelbaum KA, Evans DD, Jurish SE, Teders SJ, Rodichok LD, Barron KD. The efficacy and cost-effectiveness of minimal-therapist contact, nondrug treatments of chronic migraine and tension headache. Headache 1985a;25:214–220.
78. Blanchard EB, Appelbaum KA, Nicholson NL, Radnitz CL, Morrill B, Michultka D, Kirsch C, Hillhouse J, Dentinger MP. A controlled evaluation of the addition of cognitive therapy to a home-based biofeedback and relaxation treatment of vascular headache. Headache 1990a;30:371–376.
79. Jurish SE, Blanchard EB, Andrasik F, Teders SJ, Neff DF, Arena JG. Home versus clinic-based treatment of vascular headache. J Consult Clin Psychol 1983;51:743–751.
80. Tobin DL, Holroyd KA, Baker A, Reynolds RVC, Holms JE. Development and clinical trial of a minimal contact, cognitive-behavioral treatment for tension headache. Cognit Ther Res 1988;12:325–339.
81. Blanchard EB, McCoy GC, Musso A, Gerardi RJ, Cotch PA, Siracusa K, Andrasik F. A controlled comparison of thermal biofeedback and relaxation training in the treatment of essential hypertension: I. Short-term and long-term outcome. Behav Ther 1986;17:563–579.
82. Devineni T, Blanchard EB. A Randomized Controlled Trial of an Internet-based Treatment for Chronic Headache. Behav Res Ther 2005;43:277–292.
83. Arena JG, Dennis N, Devineni T, McClean R, Meador KJ. A pilot study of the feasibility of a telemedicine delivery system for psychophysiological treatments for vascular headache. Tele Meo J E-Health 2005;10:449–454.

See also BIOELECTRODES; ELECTROENCEPHALOGRAPHY; ELECTROGASTROGRAM; ELECTROMYOGRAPHY.

BIOHEAT TRANSFER

JONATHAN W. VALVANO
The University of Texas
Austin, Texas

INTRODUCTION

Bioheat transfer is the study of the transport of thermal energy in living systems. Because biochemical processes are temperature dependent, heat transfer plays a major role in living systems. Also, because the mass transport of blood through tissue causes a consequent thermal energy transfer, bioheat transfer methods are applicable for diagnostic and therapeutic applications involving either mass or heat transfer. This article presents the characteristics of bioheat transfer that distinguish it from nonliving systems, including the effects of blood perfusion on temperature distribution, coupling with biochemical processes, therapeutic and injury processes, and thermoregulation.

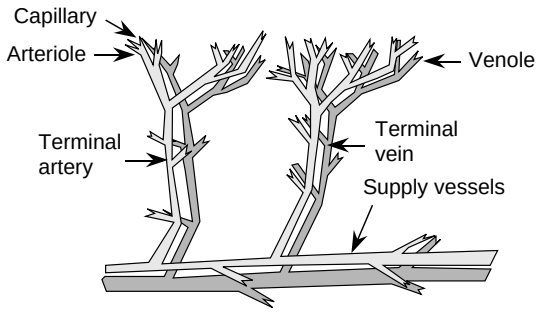


Figure 1. Countercurrent blood vessels have arterial blood flowing in the opposite direction as venous blood.

The study of bioheat transfer involves phenomena that are not found in systems that are not alive. For example, blood perfusion is considered a three-dimensional (3D) process as fluid traverses in a volumetric manner through tissues and organs via a complex network of branching vessels. Heat transfer is affected by vessel geometry, local blood flow rates, and thermal capacity of the blood (1). One factor that makes modeling blood perfusion difficult is the complex network of pairs of arteries and veins with countercurrent flow (2), as shown in Fig. 1. Arterial and venous blood temperatures may be different, and it is possible that neither is equal to the local tissue temperature. These temperatures may vary as a function of many transient physiological and physical parameters. The regulation of temperature and blood flow is quite nonlinear and has presented a major challenge to understand and model. Nevertheless, these critical processes must be accounted for in the design of many types of systems that interface with humans and animals.

Many scientists view life from either the macroscopic (systems) or the microscopic (cellular) level, but in reality one must be aware that life processes exist continuously throughout the spectrum. In order to better understand life processes at the molecular level, a significant research effort is underway associated with molecular biology. Because temperature and blood flow are critical factors, bioengineers are collaborating with molecular biologists to understand and manipulate the molecular and biochemical processes that constitute the basis of life. Research has found that the rates of nearly all physiological functions are altered 6–10%/°C (3). Similarly, heat can be added or removed during therapeutic or diagnostic procedures to produce or measure a targeted effect, based on the fact that a change in local temperature will have a large effect on rates of biochemical process rates. Thus, the measurement and control of temperature in living tissues is of great value in both the assessment of normal physiological function and the treatment of pathological states.

The study of the effects of temperature alterations on biochemical rate processes has been divided into three broad categories: hyperthermia (increased temperature), hypothermia (decreased temperature), and cryobiology

(subfreezing temperature). An extensive review of these domains has been published (4), to which the reader is referred for further details and bibliography.

Effects of Blood Perfusion on Heat Transfer

Blood perfusion through the vascular network and the local temperature distribution are interdependent. Many environmental (e.g., heat stress and hypothermia), pathophysiologic (e.g., inflammation and cancer), therapeutic (e.g., heating–cooling pads) situations create a significant temperature difference between the blood and the tissue through which it flows. The temperature difference causes convective heat transport to occur, altering the temperatures of both the blood and the tissue. Perfusion-based heat transfer interaction is critical to a number of physiological processes, such as thermoregulation and inflammation.

The convective heat transfer depends on the rate of perfusion and the vascular anatomy, which vary widely among the different tissues, organs of the body, and pathology. Diller et al. published an extensive compilation of perfusion data for many tissues and organs and for many species (5). Charney reviewed the literature on mathematical modeling of the influence of blood perfusion on bioheat transfer phenomena (6).

The rate of perfusion of blood through different tissues and organs varies over the time course of a normal day's activities, depending on factors, such as physical activity, physiological stimulus, and environmental conditions. Further, many disease processes are characterized by alterations in blood perfusion, and some therapeutic interventions result in either an increase or decrease in blood flow in a target tissue. For these reasons, it is very useful in a clinical context to know what the absolute level of blood perfusion is within a given tissue. Many thermal techniques have been developed that directly measure heat flux to predict blood perfusion by exploiting the coupling between vascular perfusion and local tissue temperature using inverse mathematical solutions.

In 1948, Pennes (7) published the seminal work describing the mathematical coupling between the mass transfer of blood perfusion and the thermal heat transfer. His work consisted of a series of experiments to measure temperature distribution as a function of radial position in the forearms of nine human subjects. A butt-junction thermocouple was passed completely through the arm via a needle inserted as a temporary track, with the two leads exiting on opposite sides of the arm. The subjects were unanesthetized so as to avoid the effects of anesthesia on blood perfusion. Following a period of normalization, the thermocouple was scanned transversely across the mediolateral axis to measure the temperature as a function of radial position within the interior of the arm. The environment in the experimental suite was kept thermally neutral during experiments. Pennes' data showed a temperature difference of 3–4° between the skin and the interior of the arm, which he attributed to the effects of metabolic heat generation and heat transfer with arterial blood perfused through the microvasculature.

Pennes proposed a model to describe the effects of metabolism and blood perfusion on the energy balance within tissue. These two effects were incorporated into the standard thermal diffusion equation, which is written in its simplified form as

$$\rho c \frac{\partial T}{\partial t} = \nabla \cdot k \nabla T + \rho_{\mathrm{bl}} c_{\mathrm{bl}} w (T_a - T) + Q_{\mathrm{met}} \qquad (1)$$

Metabolic heat generation, Q_{met}, is assumed to be homogeneously distributed throughout the tissue of interest as rate of energy deposition per unit volume. It is assumed that the blood perfusion effect is homogeneous and isotropic, and that thermal equilibration occurs in the microcirculatory capillary bed. In this scenario, blood enters capillaries at the temperature of arterial blood, T_a, where heat exchange occurs to bring the temperature to that of the surrounding tissue, T. There is assumed to be no energy transfer either before or after the blood passes through the capillaries, so that the temperature at which it enters the venous circulation is that of the local tissue. The total energy exchange between blood and tissue is directly proportional to the density, ρ_{bl}, specific heat, c_{bl}, and perfusion rate, w, of blood through the tissue. The basic principle that couples mass transfer to heat transfer is the change in sensible energy caused by the moving blood. The units of perfusion in equation 1 are volume of blood per volume of tissue per time (s^{-1}). This thermal transport model is analogous to the process of mass transport between blood and tissue, which is confined primarily to the capillary bed.

A major advantage of the Pennes model is that the added term to account for perfusion heat transfer is linear in temperature, which facilitates the solution of Eq. 1. Since the publication of this work, the Pennes model has been adapted by many researchers for the analysis of a variety of bioheat transfer phenomena. These applications vary in physiological complexity from a simple homogeneous volume of tissue to thermal regulation of the entire human body (8,9). As more scientists have evaluated the Pennes model for application in specific physiological systems, it has become increasingly clear that many of the assumptions to the model are not valid. For example, it is now well established that the significant heat transfer due to blood flow occurs in the terminal arterioles (vessels 60–300 μm in diameter) (10–17). Thermal equilibration is essentially complete for vessels <60 μm (precapillaries and capillaries). Therefore, no significant heat transfer occurs in the capillary bed; the exchange of heat occurs in the larger components of the vascular tree. The vascular morphology varies considerably among the various organs of the body, which contributes to the need for specific models for the thermal effects of blood flow (as compared to the Pennes model that incorporates no information concerning vascular geometry). It would appear as a consequence of these physiological realities that the validity of the Pennes model is questionable.

Many investigators have developed alternative models for the exchange of heat between blood and tissue. These models have accounted for the effects of vessel size, countercurrent heat exchange, as well as a combination of partial countercurrent exchange and bleed-off perfusion. All of these models provided a larger degree of rigor in the analysis, but at the compromise of greater complexity and

reduced generality. These studies also led to an increased appreciation of the necessity for a more explicit understanding of the local vascular morphology as it governs bioheat transfer, which has given rise to experimental studies to measure and characterize the 3D architecture of the vasculature in tissues and organs of interest (18). The quantitative analysis of the effects of blood perfusion on the internal temperature distribution in living tissue remains a topic of active research after one-half of a century of study (19).

THERAPEUTIC APPLICATIONS OF BIOHEAT TRANSFER

The elevation of tissue temperature into the 40–42 °C range provides some relief from pain (analgesia). In addition, wound healing can be enhanced by this modest increase in temperature. Increased temperature does not cause healing by itself, but rather it creates the improved conditions for the natural processes to heal wounds.

Hot or cold packs can be used to create therapeutic heating for injuries near the skin surface. Heating packs are effective for injuries such as sprains, muscle strains, and postoperative swelling. Elevated temperatures cause an increase in blood perfusion, supplying nutrients to the injured tissue. During the first 12–24 h after injury, cold packs will reduce perfusion, thereby reducing vascular pressure and tissue swelling. Afterward, hot packs (up to 45 °C) are applied to increase perfusion and promote healing (20).

When the injury is deeper, the therapy requires a volumetric heater, such as radio frequency (rf) electromagnetic heating, microwave frequency electromagnetic heating, and ultrasonic heating. The Industrial-Medical-Scientific (ISM) frequencies are 6.78, 13.56, 27.12, and 40.68 MHz. The ISM frequencies typically used in medical applications of microwaves are 915 MHz and 2.45 GHz. Medical ultrasonic devices operate in the 500 kHz to 10 MHz range. There are three engineering parameters to consider when designing a therapeutic device. The first parameter is the amount of local volumetric heat generation, the second parameter is the shape of the heating field, and the third parameter is the depth of penetration. Higher frequencies have shorter wavelengths, causing higher absorption and less penetration. The electrical properties of the tissue, which depend on structure and composition, strongly affect the effectiveness of volumetric heating using rf and microwave EM fields. Similarly, acoustic properties of the tissue are important in ultrasonic systems. Often the design of an effective therapeutic device hinges on proper control of the boundary conditions where energy is transferred across the transducer–tissue interface.

There is a systemic effect of local heating, controlled by the hypothalamus, involving both neuronal and hormonal signals. Local heating of organs or peripheral muscles can also cause a spinal cord mediated response. A local release of bradykinins can affect the vascular tone of the terminal arterioles (see Fig. 1), which in turn affect the vascular resistance to blood flow. In general, an increase in local temperature causes an increase in local blood flow, whereas a decrease in local temperature creates a decrease in local blood flow, but the behavior is highly complex.

Smooth muscles surrounding the 40–200 μm diameter arterioles play a dominate role in controlling local blood flow. Normal capillary pressure is ~3.3 kPa (25 Torr). When local tissues are heated, these arterioles dilate causing an increase in capillary pressure and capillary blood flow. Edema occurs when this pressure widens gaps in the capillary wall causing excess fluid to leak from the vascular to intravascular space. High capillary flow promotes healing by removing wastes, delivering nutrients, and supplying oxygen. Leukocytes (white blood cells) control the healing process by first breaking down, then removing damaged and dead tissue.

The volumetric heating created by electromagnetic fields in governed by the electrical conductivity, σ ($S \cdot m^{-1}$), the imaginary part of the electrical permittivity, ε'' ($F \cdot m^{-1}$), and the magnitude of the local electric field, $|\boldsymbol{E}|$ ($V \cdot m^{-1}$):

$$q''' = (\sigma + \omega\varepsilon'')|\boldsymbol{E}|^2 \tag{2}$$

where ω is the angular frequency of the field in rad s^{-1}. Direct heating from the magnetic fields in medical applications is usually neglected. Magnetic fields can be used to heat tissue, but because of Faraday's law of induction, the time-varying magnetic field will induce an electric field, and it is this electric field that heats the tissue. A comprehensive review of electromagnetic heating can be found in Roussy and Pearce (21). Tables of electrical and acoustic properties can be found in Diller (5).

THERMOREGULATION

Thermoregulation is an elaborate control system, used by mammals, to maintain internal body temperatures near a physiological set point under a large spectrum of environmental conditions and metabolic rate activities. Even though there have been many years of research, much remains unknown about the human thermoregulatory system. Therefore, active investigation continues. Heat transfer due to conduction, convective heat transfer via the blood flow, local generation of thermal energy, and thermal boundary conditions comprise the major components of thermoregulation. Once these individual mechanisms are understood, they can be combined to create mathematical models to simulate and predict thermoregulatory behavior. The mathematical models are used to design systems to interact thermally with the human body (such as a space suit) without compromising the health and safety of the subject.

The prevailing theory is that the main objective of a human thermoregulation system is to maintain the body core temperature at a constant value consistent with that required for normal physiological functions, regardless of the environmental conditions. An alternative theory, suggested by Chappuis et al. (22) and Webb (23), is that the goal of the human regulation system is to maintain the body's energy balance. In this theory, tissue temperatures are a result, rather than a cause, of the regulation process.

Nunneley (24,25) showed that temperature and internal energy storage of the human body vary with time of day, metabolic activity, and individuality of the human. To maintain body core temperature, the thermoregulatory system incorporates a number of energy production and dissipation mechanisms, many of them controlled by feedback from other body parameters. Examples of such feedback control are that for sweating, shivering, and varying blood flow.

Ganong (26) showed that the main control center for feedback mechanisms is located in the hypothalamus of the brain, where reflex responses operate to maintain the body temperature within its narrow range. The signals that activate the hypothalamic temperature regulating centers come largely from two sources: the temperature-sensitive cells in the anterior hypothalamus and cutaneous temperature receptors. The cells in the anterior hypothalamus sense the temperature of the body core or, specifically, the temperature of the arterial blood that passes through the head.

The theory of energy regulation is based on the demonstration of the existence of temperature sensors at several levels in the skin enabling the sensing of heat flow within and from the body. Evidence has also shown neurological sensing of thermal gradients, of which changes relate to the thermal regulating responses. Therefore, the hypothesis behind the theory of energy content regulation based on Webb's experimental observations is "Heat (energy) regulation achieves heat (energy) balance over a wide range of heat (energy) loads. Heat flow to or from the body is sensed, and physiological responses defend the body heat (energy) content. Heat (energy) content varies over a range that is related to heat (energy) load. Changes in body heat (energy) content drive deep body temperatures" (23). Energy regulation involves constantly changing metabolic energy production and the adjustment of heat losses to maintain a system in equilibrium. This mechanism is opposed to temperature regulation where adjustments are required to maintain body temperature.

The thermal energy balance over time within the human body combines the heat added by internal production minus the heat lost by various heat-transfer processes.

$$\Delta E = M - (W + Q_{\text{conv}} + Q_{\text{cond}} + Q_{\text{rad}} + Q_{\text{evap}} + Q_{\text{resp}}) \qquad (3)$$

where ΔE is the rate of energy storage in the body (W), M is the metabolic energy production (W), W is the external work (W), Q_{conv} is the surface heat loss by convection (W), Q_{cond} is the surface heat loss by conduction (W), Q_{rad} is the surface heat loss by radiation (W), Q_{evap} is the surface heat loss by evaporation (W), and Q_{resp} is the respiratory heat loss (W).

The human body produces energy, exchanges heat with the environment, and loses heat by evaporation of body fluids. Thermal energy is produced by metabolism, a biochemical process occurring in cells has adenosine triphosphate (ATP) is combined with oxygen to produce the various life functions. Fulcher (27) defined the basal metabolic rate as "the minimal metabolism measured at a temperature of thermal neutrality in a resting homeotherm with normal body temperature several hours after a meal and not immediately after hypothermia". Energy is also produced at an increased rate due to muscle activity, including physical exercise and shivering, and by food intake. Therefore, the total energy production in the body is determined by the energy needed for basic body processes plus any external work. Since the body operates with $<100\%$ efficiency only a fraction of the metabolic rate is applied to work. The remainder shows up as heat. The mechanical efficiency associated with metabolic energy utilization is zero for most activities except when the person is performing external mechanical work, such as in walking upstairs, lifting something to a higher level, or cycling on an exercise machine. When external work is dissipated into heat in the human body, the mechanical efficiency is negative. An example of negative mechanical efficiency is walking downstairs.

Convection, radiation, conduction, and evaporation of sweat at the skin surface allow heat to be dissipated from the body. There is also heat transfer, especially when the environmental air temperature is extremely high or low, through the respiratory tract and lungs. Storage of energy takes place whenever there is an imbalance of production and dissipation mechanisms. In many instances, such as astronauts in space suits or military personnel in chemical defense garments, energy storage is forced due to the lack of appropriate heat exchange with the environment.

The human thermoregulatory system is quite complicated and behaves mathematically in a highly nonlinear manner. It contains multiple sensors, multiple feedback loops, and multiple outputs. The control mechanisms to release excess energy include the production of sweat, and vasodilatation of the blood vessels in the skin. Conversely, to conserve energy there can be shivering of the muscles, and vasoconstriction of blood vessels, which engage in the transportation of heat to the surface of the body.

Heat transfer within the body is due to the internal conductance that governs the flow of heat from the core, through the tissue, to the surface. This component of heat transfer is governed by peripheral blood flow, the core-skin temperature gradient, and the conductivity of the body tissue. Blood flow provides the majority of the peripheral conductance where there is convection between blood and tissue and countercurrent heat exchange between the arteries and the veins. Blood flow is controlled according to metabolic needs of the body as well as the need to maintain the appropriate core temperature. When the core becomes too hot, the blood vessels in the skin dilate to allow increased blood flow to the surface of the skin. Then, the environment cools the blood and the cooler blood returns to the core. Increased blood flow to the skin surface increases extravascular pressure enabling greater sweat production, again adding to the cooling process. In contrast, when the core becomes too cold, blood flow to the skin is constricted to conserve the body's internal energy.

Sweating is centrally controlled by the hypothalamus. When the body senses an increase temperature the hypothalamus increases nerve impulses to the sweat glands. Shivering, on the other hand, is an involuntary response of the skeletal muscles when passive body cooling exceeds metabolic energy production.

This section briefly introduced the concepts of thermoregulation. For a quantitative analysis of this topic, see Wissler (8,9).

THERMAL INJURY

Thermal injury is defined as irreversible changes to living tissue caused by temperature. Injury can occur when the tissue temperature exceeds the range between which normal life processes exist. Both high and low temperature can cause irreversible changes to biomolecules, resulting in injury. Common examples are burns and frostbite. Recently, it has been discovered that under some kinds of moderate thermal stress that is subthreshold to injury, cells produce molecules that render temporary protection against levels of many types of stress (thermal, mechanical, chemical, etc.) that would normally cause injury. These protective molecules are called heat shock proteins, and they are the subject of widespread investigation to identify the kinetics of their expression and function. It is an effort to develop applications in which they may be induced either before or even after a traumatic event.

The most commonly encountered type of thermal injury is the burn. Accidental burns are encountered most frequently in domestic and industrial settings as well as many other venues of activity. Most burns result from the propagation of heat inward into tissues as a result of contact at the surface (skin) with a hot solid, liquid, or vapor. One exception is electrical burns in which the tissue temperature is elevated owing to I^2R dissipation of electric energy when a voltage is applied. In this case, the primary source of heating is internal since the impedance of muscle is higher than of skin and fat.

It is generally assumed that thermal burns can be modeled as a simple Arrhenius rate process such that

$$\Omega(t) = \int_0^t A e^{\Delta E/RT(\tau)} d\tau \tag{4}$$

where Ω is a dimensionless damage parameter (e.g., $\Omega = 1$ means first degree burn), A is a frequency factor (s^{-1}), ΔE is the activation energy in $J \cdot mol^{-1}$, R is the universal gas constant ($8.314\,J \cdot mol^{-1} \cdot K^{-1}$), and T is the tissue temperature (K). The constants A and ΔE are tissue parameters, and $T(\tau)$ is the time history of the tissue temperature (28).

This model was first posed for predicting the severity of a burn as a function of the temperature and time of exposure at the skin surface by Moritz and Henriques (29) shortly after World War II. They also performed experiments to determine threshold conditions for eliciting first and second degree burns in humans and applied this data to determine values for the scaling constant and activation energy in their Arrhenius model. Their experiments were conducted at temperatures between 44 and 70°C and exposure times between 1 and 25,000 s. The model parameter values are $A = 3.1 \times 10^{98}\,s^{-1}$ and $\Delta E = 6.28 \times 10^5\,kJ \cdot mol^{-1}$. Over the ensuing 50 years many subsequent investigators have studied this process with mathematical models and experimental investigations (30–40). Although a considerable body of literature has been accrued, there is by no means a consensus on how to accurately predict the occurrence of thermal injury over the wide range of conditions that cause burns.

SUBZERO EFFECTS

One application of subzero temperatures is the long-term preservation of biologic tissue. Therapeutic devices based on subzero temperatures can be used to destroy cancerous cells or remove necrotic tissue. The rate of biochemical processes is governed by local temperature. Lowering the temperature has the effect of reducing reaction rates, and at sufficiently low temperatures, a state of suspended animation can be achieved. Because of the water component of physiological fluids, temperatures low enough to affect suspended animation normally result in freezing. The freezing of native biomaterials is nearly always lethal to the affected tissue upon thawing. The formation of ice has two damaging effects. The first effect is mechanical as intracellular ice crystals physically damage cell structures. The second and more lethal effect is osmotic. The local concentration of ions, such as Na^+, K^+, and Cl^-, are critical for sustaining life, and a high concentration of these ions is produced as liquid water freezes into ice. The effected injury can be used to benefit in cryosurgery for the purpose of destroying a target tissue, such as cancer. Alternatively, the tissue can be modified prior to freezing by the introduction of a chemical cryoprotective agent (CPA) to afford protection from freeze–thaw injury. The CPA either protects against the injurious effects of ice formation or blocks the formation of ice so that a glassy state results, which is called vitrification. Organ transplantation, blood banks, and animal husbandry are three applications that require the successful long-term cryopreservation of biologic tissue. The response of living biomaterials to freezing and thawing is intimately tied to the thermal history during processing, especially at subzero temperatures. Thus, bioheat transfer analysis has played a key role in the design and development of effective cryopreservation techniques. Polge was first to report the successful use of glycerol to freeze fowl sperm >55 years ago (41). Successes were reported in succession for other types of tissues having rather simple cell structures, such as erythrocytes, gametes and various cells obtained from primary cultures (42–44). Most of these cryopreservation techniques were derived via largely empirical methods, and starting in the 1970s it came to be realized that the cryopreservation of more complex systems, such as multicellular tissues and whole organs require a more rigorous scientific understanding of the mechanisms of the governing biophysical processes and cellular response to freezing and thawing. Since that time engineers have made significant contributions to the developing science of cryobiology, not the least of which has been to identify some of the key biophysical problems to be solved (45,46).

MEASUREMENT OF THERMAL CONDUCTIVITY AND THERMAL DIFFUSIVITY

While the other sections in this article presented brief overviews of the various disciplines within the field of

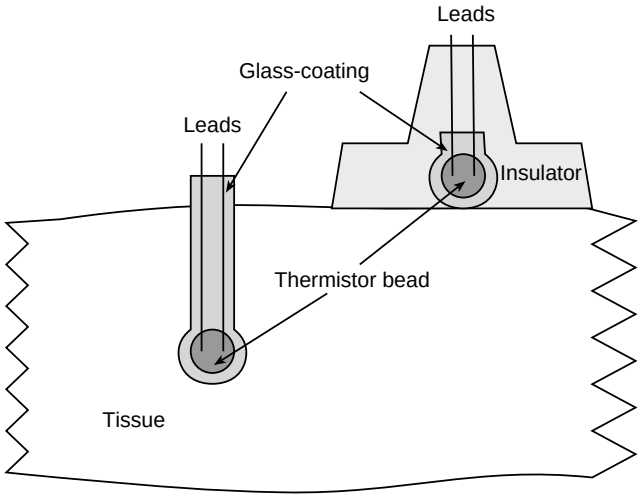


Figure 2. A glass-coated thermistor is placed into or on the surface of the tissue of interest.

bioheat transfer, this section will present in detail a specific measurement technique. In particular, this section presents an instrument used to measure thermal properties in tissue. The section begins with definitions of thermal properties, overviews of the technique, then develops the heat-transfer equations that form the basis of the instrument. Finally, calibration methods and error analyses are presented.

Definitions of Thermal Properties

Thermal conductivity (k) is the ability of a material to transport heat in the steady state. In one dimension, the total heat (Q) transported across a flat surface of area A and thickness Δx is related to the temperature gradient across the surface (ΔT) and the thermal conductivity of the material.

$$Q = -kA\frac{\Delta T}{\Delta x} \tag{5}$$

Thermal diffusivity (α) is the ability of a material to conduct heat in the transient state. Thermal properties of conductivity and diffusivity are related. The quotient of conductivity divided by diffusivity equals density times specific heat.

$$\frac{k}{\alpha} = \rho c \tag{6}$$

Diffusivity is often defined in the partial differential equation used to describe transient heat transfer. Assuming homogeneous thermal properties, the Fourier conduction equation in one dimension is

$$\frac{\partial^2 T}{\partial x^2} = \frac{1}{\alpha}\frac{\partial T}{\partial t} \tag{7}$$

Measurement Technique

The technique involves inserting a thermistor into the tissue of interest or placing it on the tissue surface, as shown in Fig. 2. Thermometrics P60DA102M and Fenwal 121-102EAJ-Q01 are glass probe thermistors that make excellent transducers (shown on the left of Fig. 2). The diameter of these thermistors is ~0.15 cm. The glass-coated spherical probes provide a large bead size and a rugged, stable transducer. The Thermometrics BR55KA102M and Fenwal 112-102EAJ-B01 bead thermistors also provide excellent results (shown on the right of Fig. 2).

If the tissue is living, the properties measured are called effective thermal conductivity, k_{eff}, and effective thermal diffusivity, α_{eff}. Effective thermal properties include the contribution to heat transfer due to intrinsic conduction added to the contribution caused by the transport of blood through the tissue.

In the constant temperature heating technique (47–52), the instrument first measures the baseline tissue temperature, T_s. Then, an electronic feedback circuit applies a variable voltage, $V_0(t)$, in order to maintain the average thermistor temperature at a predefined constant, T_h. The electrical circuit used to implement the constant temperature heating technique is shown in Fig. 3. Three high quality, gold-plated, electromagnetic relays are used to switch the thermistor (R_s) between "heat" and "sense" mode. Figure 3 shows the position of the three relays in "heat" mode. Initially, the instrument places the circuit in "sense" mode with the three relays in

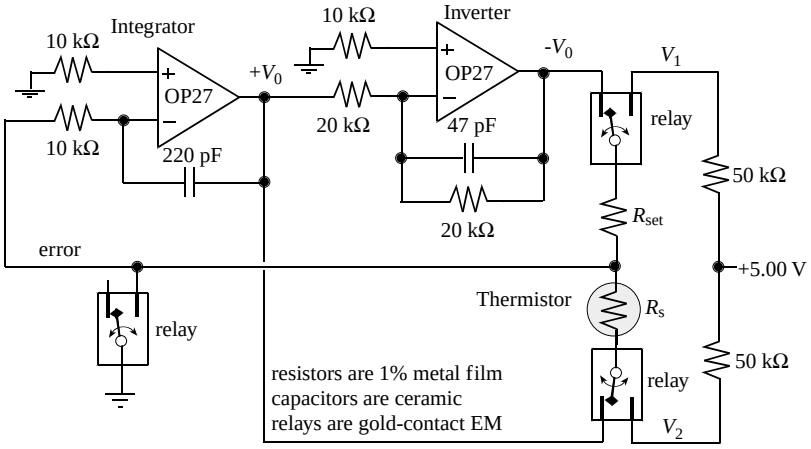


Figure 3. Instrumentation used for the constant temperature heating technique.

the opposite position as shown in Fig. 3. A precision +5.00 V reference (PMI REF02) supplies voltage to the four-resistor bridge, formed by the two 50 kΩ, R_{set}, and R_s resistors. The voltage difference $V_2 - V_1$ is fed to a differential amplifier, passed through a low pass filter, then fed to a 12-bit ADC.

Fundamental Equations

Resistance calibration is performed to determine the relationship between the ADC sample and the unknown R_s. Next, temperature calibration is performed by placing the thermistor adjacent to an accurate temperature monitor and placing the combination in a temperature- controlled waterbath. The thermistor resistance varies nonlinearly with its temperature. For small temperature ranges equation 8 can be used for temperature calibration.

$$R_s = R_{\text{O}}\, \text{e}^{\beta/(T_s+273.15)} \tag{8}$$

where T_s is the temperature in degrees Celsius, and R_s is the thermistor resistance in ohms.

In heat mode, the integrator–inverter circuit varies the voltage across the thermistor until the thermistor resistance, R_s, matches the fixed resistor, R_{set}. It takes just a few milliseconds for the electrical control circuit to stabilize. Once stable, R_s is equal to R_{set}, meaning the volume average thermistor temperature is equal to a constant. The instrument uses a calibration temperature versus resistance curve to determine the heated temperature T_h from the fixed resistor R_{set}. The power applied to the thermistor, P, is calculated from $(V_0)^2/R_{\text{set}}$. The applied thermistor power includes a steady state and a transient term:

$$P(t) = A + Bt^{-1/2} \tag{9}$$

In order to measure thermal conductivity, thermal diffusivity, and tissue perfusion the relationship between applied thermistor power, P, and resulting thermistor temperature rise, $\Delta T(t) = T_h - T_s$, must be known. In the constant temperature method, ΔT is constant. The thermistor bead is treated as a sphere of radius a embedded in a homogeneous medium. Since all media are considered to have constant parameters with respect to time and space, the initial temperature will be uniform when no power is supplied to the probe.

$$T_b = T_m = T_s = T_a + \frac{Q_{\text{met}}}{w\rho_{\text{bl}}c_{\text{bl}}} \qquad \text{at} \quad t = 0 \tag{10}$$

Let V be the temperature rise above baseline, $V = T - T_s$. Both the thermistor bead temperature rise (V_b) and the tissue temperature rise (V_m) are initially zero.

$$V_b = V_m = 0 \qquad \text{at} \quad t = 0 \tag{11}$$

To solve this coupled thermistor–tissue system, equation 7 is written in spherical coordinates and the applied power is deposited into the thermistor, while the perfusion heat sink is added to the tissue, equation 1. Assuming the venous blood temperature equilibrates with the tissue temperature and that the metabolic heat is uniform in time and space, the Pennes' bioheat transfer equation in spherical coordinates is given by

$$\rho_b c_b \frac{\partial V_b}{\partial t} = k_b \frac{1}{r^2}\frac{\partial}{\partial r}\left(r^2 \frac{\partial V_b}{\partial r}\right) + \frac{A + Bt^{-1/2}}{4/3\pi a^3} \qquad r < a \tag{12}$$

$$\rho_m c_m \frac{\partial V_m}{\partial t} = k_m \frac{1}{r^2}\frac{\partial}{\partial r}\left(r^2 \frac{\partial V_m}{\partial r}\right) - w\rho_{bl}c_{bl}V_m \qquad r > a \tag{13}$$

where w is the tissue perfusion (s^{-1}). Perfect thermal contact is assumed between the finite-sized spherical thermistor and the infinite homogeneous perfused tissue. At the interface between the bead and the tissue, continuity of thermal flux and temperature leads to the following boundary conditions:

$$V_b = V_m \qquad \text{at } r = a \tag{14}$$

$$k_b \frac{\partial V_b}{\partial r} = k_m \frac{\partial V_m}{\partial r} \qquad \text{at } r = a \tag{15}$$

The other boundary conditions are necessary at positions $r \to 0$ and $r \to$ infinity. Since no heat is gained or lost at the center of the thermistor:

$$k_b \frac{\partial V_b}{\partial r} = 0 \qquad \text{as } r \to 0 \tag{16}$$

Because the thermistor power is finite and the tissue is infinite, the tissue temperature rise at infinity goes to zero:

$$V_m \to 0 \qquad \text{as } r \to \text{infinity} \tag{17}$$

It is this last initial condition that allows the Laplace transform to be used to solve the coupled partial differential equations. The Laplace transform converts the partial differential equations into ordinary differential equations that are independent of time t. The steady-state solution allows for the determination of thermal conductivity and perfusion (49).

$$V_b(r) = \frac{A}{4\pi a\, k_b}\left(\frac{k_b}{k_m(1+\sqrt{z})} + \frac{1}{2}\left[1 - \left(\frac{r}{a}\right)^2\right]\right) \tag{18}$$

$$V_m(r) = \frac{A}{4\pi r\, k_m}\left(\frac{e^{(1-r/a)\sqrt{z}}}{1+\sqrt{z}}\right) \tag{19}$$

where z is a dimensionless Pennes' model perfusion term ($w\rho_{\text{bl}}\, c_{\text{bl}}\, a^2/k_m$). The measured thermistor response, ΔT, is assumed be the simple volume average of the thermistor temperature:

$$\Delta T = \frac{\int_0^a V_b(r) 4\pi r^2 dr}{4/3\pi a^3} \tag{20}$$

Inserting equation 18 into Eq. 20 yields the relationship used to measure thermal conductivity assuming no perfusion (49).

$$k_m = \frac{1}{\frac{4\pi a\, \Delta T}{A} - \frac{0.2}{k_b}} \tag{21}$$

A similar equation allows the measurement of thermal diffusivity from the transient response, again assuming no

perfusion (49).

$$\alpha_m = \left(\frac{a}{\sqrt{\pi}\,\frac{B}{A}\left(1 + 0.2\,\frac{k_m}{k_b}\right)} \right)^2 \tag{22}$$

Calibration Equations

The first calibration determines relationship between the ADC sample and the thermistor resistance when in sense mode. For this calibration, precision resistors are connected in place of the thermistor, and the computer-based instrument is used to sample the ADC in sense mode. A simple linear equation works well for converting ADC samples to measured resistance. In this procedure, the device acts like a standard ohmmeter.

The second calibration determines the relationship between thermistor temperature and its resistance. The instrument measures resistance, and a precision thermometer determines true temperature. Equation 23 yields an accurate fit over a wide range of temperature:

$$T = \frac{1}{H_0 + H_1 \ln(R) + H_3 [\ln(R)]^3} - 273.15 \tag{23}$$

where T is in degrees Celsius. Temperature resistance data are fit to Eq. 23 using nonlinear regression to determine the calibration coefficients H_0, H_1, and H_3.

The applied power, $P(t)$, is measured during a 30 s transient while in heat mode. Nonlinear regression is used to calculate the steady-state and transient terms in equation 9. Figure 4 shows some typical responses. The steady-state response (time equals infinity) is a measure of the thermal conductivity. The transient response (slope) indicated the thermal diffusivity.

The third calibration maps measured power to thermal properties while operating in heat mode. Rather than using the actual probe radius (a) and probe thermal conductivity (k_b), as shown in Eqs. 21 and 22, the following empirical equations are used to calculate thermal properties.

$$k_m = \frac{1}{(c_1\,\Delta T/A) + c_2} \tag{24}$$

$$\alpha_m = \left(\frac{c_3}{B/A(1 + k_m/c_4)} \right)^2 \tag{25}$$

The coefficients c_1, c_2, c_3, and c_4 are determined by operating the probe in two materials of known thermal properties. Typically, agar-gelled water and glycerol are used as thermal standards. This empirical calibration is performed at the same temperatures at which the thermal property measurements will be performed.

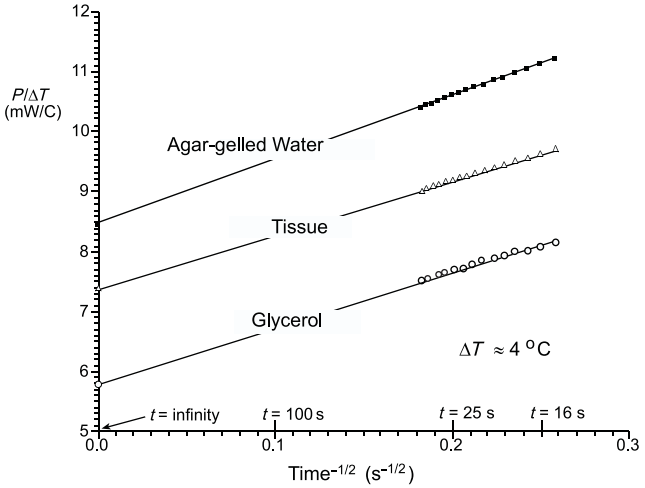


Figure 4. Typical $P/\Delta T$ versus $t^{-1/2}$ data for the constant temperature heating technique. The agar-gelled water and glycerol curves are used for empirical calibration.

Error Analysis

It is assumed that the baseline tissue temperature, T_0, is constant during the 30 s transient. Patel has shown that if the temperature drift, dT_0/dt, is $>0.002\,^{\circ}\mathrm{C}\cdot\mathrm{s}^{-1}$, then significant errors will occur (52). The electronic feedback circuit forces T_h to a constant. Thus, if T_0 is constant then ΔT does not vary during the 30 s transient.

The time of heating can vary from 10 to 60 s. Shorter heating times are better for small tissue samples and for situations where there is baseline tissue temperature drift. Another advantage of shorter heating times is the reduction in the total time required to make one measurement. Longer heating times increase the measurement volume and reduce the effect of imperfect thermistor–tissue coupling. Typically, shorter heating times are used *in vivo* because it allows more measurements to be taken over the same time period. On the other hand, longer heating times are used *in vitro* because accuracy is more important than measurement speed.

Thermal probes must be constructed in order to measure thermal properties. The two important factors for the thermal probe are thermal contact and transducer sensitivity. The shape of the probe should be chosen in order to minimize trauma during insertion. Any boundary layer between the thermistor and the tissue of interest will cause a significant measurement error. The second factor is transducer sensitivity that is the slope of the thermistor voltage versus tissue thermal conductivity. Equation 21 shows for a fixed ΔT k_m and k_b the thermistor power (A) increases linearly with probe size (a). Therefore larger probes are more sensitive to thermal conductivity. For large tissue samples, multiple thermistors can be wired in parallel, so they act electrically and thermally as one large device. There are two advantages to using multiple thermistors. The effective radius, $a = c_1/4\pi$, is increased from $\sim$0.08 cm for a typical single P60DA102M probe to $\sim$0.5 cm for a configuration of three P60DA102M thermistors. The second advantage is that the three thermistors are close enough to each other that the tissue between the probes will be heated by all three thermistors. This cooperative heating tends to increase the effective measurement volume and reduce the probe/tissue contact error. Good mechanical–thermal contact is critical. The probes are calibrated after they are constructed, so that the thermistor geometry is incorporated into the coefficients

c_1, c_2, c_3, and c_4. The same waterbath, and probe configuration should be used during the calibration and during the tissue measurements.

Calibration is a critical factor when using an empirical technique. For temperatures $<0\,^{\circ}$C, ice and ethylene glycol are used as thermal standards. For temperatures between 0 and 15 °C, agar-gelled water and ethylene glycol can be used as thermal standards. For temperatures between 15 and 75 °C, agar-gelled water and glycerol were used. To prevent convection, 1 g of agar/100 mL of water should be added. A mixture of water and glycerol can be used to estimate the accuracy of the technique. The mass fraction, m, can be used to determine the true thermal properties of the mixture (53,54). The ability to determine measurement accuracy is critical for the acceptance of new technology. These two equations provide for the capability to create reference materials of known thermal properties, which can be used to experimentally determine measurement accuracy.

$$k_m = m\,k_g + (1-m)k_w + 1.4\,m(m-1)(k_w - k_g - 2) - 0.014\,m(m-1)(T - 20\,^{\circ}\mathrm{C}) \quad (26)$$

$$\alpha_m = m\,\alpha_g + (1-m)\alpha_w \quad (27)$$

where T is in degrees Celsius. Self-heat thermistors have also been successfully used to measure the convective heat transfer coefficient on the endocardial surface of the heart (55,56).

ADDITIONAL STUDIES

In this article, the general concepts of bioheat transfer were introduced, and a detailed design and analysis of an instrument that measures thermal properties was presented. Although out of print, the 1985 book *Heat Transfer in Medicine and Biology*, is a wonderful collection of detailed works that address a wide spectrum of topics in bioheat transfer. The book *Optical-Thermal Response of Laser Irradiated Tissue* covers the issues involved in high temperature effect such as tissue damage and thermal ablation. Valvano's chapter titled Temperature Measurements, in *Advances In Heat Transfer: Bioengineering Heat Transfer*, covers many practical issues involved in measuring temperature in the biomedical setting. An in depth treatment of bioheat transfer topics can be found in the 2005 edition of *CRC Handbook of Heat Transfer*. This reference has excellent treatments of thermoregulation and low temperature effects.

BIBLIOGRAPHY

Cited References

1. Baish JW, Ayyaswamy PS, Foster KR. Heat transport mechanisms in vascular tissues: a model comparison. J Biomech Eng 1986;108:324–331.
2. Baish JW. Heat transport by countercurrent blood vessels in the presence of an arbitrary temperature gradient. J Biomech Eng 1990;112:207–211.
3. Johnston KA, Bennett AF, editors. Animals and Temperature: Phenotypic and Evolutionary Adaptation. Cambridge: Cambridge University Press; 1996.
4. Diller KR. Modeling of bioheat transfer processes at high and low temperatures. Adv Heat Trans 1992;22:157–357.
5. Diller KR, Valvano JW, Pearce JA. Bioheat Transfer. In: Kneith F, editor. CRC Handbook of Heat Transfer, 2nd ed. 2005.
6. Charney CK. Mathematical models of bioheat transfer. Adv Heat Trans 1992;22:19–155.
7. Pennes HH. Analysis of Tissue and Arterial Blood Temperature in the Resting Human Forearm. J Appl Phys 1948;1:93–102.
8. Wissler E. A review of human thermal models. In: Morrison MB, editor. Environmental Ergonomics. New York: Taylor and Francis; 1988. p 267–285.
9. Wissler EH. Mathematical simulation of human thermal behavior using whole-body models. In: Shitzer A, Eberhart RC, editors. Heat Transfer in Medicine and Biology. Vol. 1. New York: Plenum Press; 1985. p 325–373.
10. Chato JC. Heat transfer to blood vessels. J Biomech Eng 1980;102:110–118.
11. Chen MM, Holmes KR. Microvascular contributions in tissue heat transfer. Annals NY Acad Sci 1980;335:137–150.
12. Weinbaum S, Jiji L, Lemons DE. Theory and Experiment for the Effect of Vascular Temperature on Surface Tissue Heat Transfer—Part 1: Anatomical Foundation and Model Conceptualization. ASME J Biomech Eng 1984;106:246–251.
13. Weinbaum S, Jiji L, Lemons DE. Theory and Experiment for the Effect of Vascular Temperature on Surface Tissue Heat Transfer—Part 2: Model Formulation and Solution. ASME J Biomech Eng 1984;106:331–341.
14. Weinbaum S, Jiji L. A New Simplified Bioheat Equation for the Effect of Blood Flow on Average Tissue Temperature. J of Biomech Eng 1985;107:131–139.
15. Charny CK, Weinbaum S, Levin RL. An Evaluation of the Weinbaum-Jiji Bioheat Equation for Normal and Hyperthermic Conditions. ASME J Biomech Eng 1990;112:80–87.
16. Xu LX, Chen MM, Holmes KR, Arkin H. The Evaluation of the Pennes, the Chen-Holmes, the Weinbaum-Jiji Bioheat Transfer Models in the Pig Kidney Cortex. ASME WAM Proc HDT 1991;189:15–21.
17. Arkin H, Xu LX, Holmes KR. Recent Developments in Modeling Heat Transfer in Blood Perfused Tissues. IEEE Trans Biomed Eng 1994;41(2):97–107.
18. Wissler EH. Pennes' 1948 paper revisited. J Appl Physiol 1998;85:35–41.
19. Pennes HH. Analysis of tissue and arterial blood temperatures in the resting forearm. J Appl Physiol 1948;1:93–122 (republished for fiftieth anniversary issue of J Appl Physiol 1998;85:5–34).
20. Scully RM, Barnes MR. Physical Therapy. Philadelphia: J.B. Lippincott Co.; 1989.
21. Roussy G, Pearce JA. Foundations And Industrial Applications Of Microwaves Physical And Chemical Processes. New York: John Wiley & Sons, Inc.; 1995.
22. Chappuis P et al. Heat storage regulation in exercise during thermal transients. J Appl Physiol 1976;40:384–392.
23. Webb P. The physiology of heat regulation. Am J Physiol 1995;268:R838–R850.
24. Nunneley SA. Water cooled garments: a review. Space Life Sci 1970;2:335–360.
25. Nunneley SA. Physiological response of women to thermal stress: A review. Med Sci Sports 1978;10:250–255.
26. Ganong WF. Review of Medical Physiology. 16th ed. Norwalk (CT): Appleton and Lange; 1993.
27. Fulcher CWG. Control of a liquid cooling garment for extravehicular astronauts by cutaneous and external auditory meatus temperatures, Ph.D. dissertation, Houston (TX): University of Houston; 1970.

28. Thomsen S. Mapping of thermal injury in biologic tissues using quantitative pathologic techniques. Proc SPIE 1999;3594-09:822–897.
29. Moritz AR, Henriques FC. Studies of Thermal Injury. II. The Relative Importance of Time and Surface Temperature in the Causation of Cutaneous Burns. Am J Path 1947;23:695–720.
30. Büttner K. Effects of extreme heat and cold on human skin. I. analysis of temperature changes caused by different kinds of heat application. J Appl Physiol 1951;3:691–702.
31. Büttner K. Effects of extreme heat and cold on human skin. II. surface temperature, pain and heat conductivity in experiments with radiant heat. J Appl Physiol 1951;3: 691–702.
32. Stoll AM. A computer solution for determination of thermal tissue damage integrals from experimental data. I R E Trans Med Electron 1960;7:355–358.
33. Stoll AM, Chianta MA. Burn production and prevention in convective and radiant heat transfer. Aerospace Med 1968;39:1232–1238.
34. Stoll AM, Green LC. Relationship between pain and tissue damage due to thermal radiation. J Appl Physiol 1959;14:373–382.
35. Ross DC, Diller KR. An experimental investigation of burn injury in living tissue. J Heat Trans 1976;98:292–296.
36. Lawrence JC, Bull JP. Thermal conditions which cause skin burns. J Inst Mech Eng Eng Med 1976;5:61–63.
37. Takata AN. Development of criterion for skin burns. Aerospace Med 1974;45:634–637.
38. Welch AJ, Polhamus GD. Measurement and prediction of thermal injury in the retina of Rhesus monkey. IEEE Trans Biomed Eng 1984;BME-31:633–644.
39. Thomsen S, Pearce JA, Cheong WF. Changes in birefringence as makers of thermal damage in tissues. IEEE Trans Biomed Eng 1989;BME-36:1174–1179.
40. Pearce JA, Thomsen S. Kinetic models of tissue fusion processes. Proc SPIE Laser Tissue Int III 1992;1643.
41. Polge C, Smith AU, Parkes AS. Revival of spermatozoa after vitrificaiton and dehydration at low temperatures. Nature 1949;164:666.
42. Lovelock JE The mechanism of the protective action of glycerol against haemolysis by freezing and thawing. Biochim Biophys Acta 1953;11:28–36.
43. Strumia MM, Clawell LS, Strumia PV. The preservation of blood for transfusion. J Lab Clin Med 1960;56:576–593.
44. Whittingham DG, Leibo SP, Mazur P. Survival of mouse embryos frozen to −196°C and −296°C. Science 1972;178: 411–414.
45. McGrath JJ, Diller KR, editors. Low Temperature Biotechnology: Emerging Applications and Engineering Contributions. New York: ASME; 1988. p 1–380.
46. Diller KR, Ryan TP. Heat transfer in living systems: current opportunities. J Heat Trans 1998;120:810–829.
47. Bowman HF. Estimation of Tissue Blood Flow. In: Shitzer, Eberhart, editors. Heat Transfer in Medicine and Biology. New York: Plenum; 1985. p 193–230.
48. Chato JC. Measurement of Thermal Properties of Biological Materials. In: Shitzer, Eberhart, editors. Heat Transfer in Medicine and Biology. New York: Plenum; 1985. p 167–192.
49. Valvano JW, et al. The simultaneous measurement of thermal conductivity, thermal diffusivity and perfusion in small volumes of tissue. J Biomech Eng 1984;106:192–197.
50. Valvano JW, et al. Thermal conductivity and diffusivity of biomaterials measured with self-heated thermistors. Intern J Thermophys 1985;6:301–311.
51. Valvano JW, Chitsabesan B. Thermal conductivity and diffusivity of arterial wall and atherosclerotic plaque. Lasers Life Sci 1987;1:219–229.
52. Patel PA, et al. A self-heated thermistor technique to measure effective thermal properties from the tissue surface. J Biomech Eng 1987;109:330–335.
53. Rastorguev YL, Ganiev YA. Thermal conductivity of aqueous solutions or organic materials. Russ J Phys Chem 1966;40: 869–871.
54. Touloukian YS, et al. Thermophysical Properties of Matter: Thermal Conductivity. Vol. 3. New York: IFI/Plenum; 1970. p 120, 209.
55. dos Santos I, et al. An instrument to measure the heat convection coefficient on the endocardial surface. Physiol Measur 2003;24:321–335.
56. dos Santos I, et al. In vivo measurements of heat transfer on the endocardial surface. Physiol Measur 2003;24:793–804.

Reading List

Welch AJ, van Gemert M, editors. Optical-Thermal Response of Laser Irradiated Tissue. New York: Plenum Press; 1995.
Valvano JW. Temperature Measurements. In: Advances In Heat Transfer: Bioengineering Heat Transfer. Vol. 22. New York: Academic Press; 1992. p 359–436.
Roussy G, Pearce JA. Foundations And Industrial Applications of Microwaves Physical And Chemical Processes. New York: John Wiley & Sons, Inc.; 1995.
Shitzer, Eberhart, editors. Heat Transfer in Medicine and Biology. New York: Plenum; 1985.
Kreith F, editor. CRC Handbook of Heat Transfer. 2nd ed. 2005.

See also CYSTIC FIBROSIS SWEAT TEST; HYPERTHERMIA, SYSTEMIC; TEMPERATURE MONITORING; THERMOMETRY.

BIOIMPEDANCE IN CARDIOVASCULAR MEDICINE

DOUGLAS A. HETTRICK
TODD M. ZIELINSKI
Medtronic, Inc.
Minneapolis, Minnesota

INTRODUCTION

Historical Context

Electrical impedance measurements have been applied to the study of biologic systems for nearly 200 years. Indeed, the history of continuously flowing electricity began with Luigi Galvani's famous experiments on bioelectricity at the University of Bologna (1,2). It was not until the 1870s however, that Hermann Müller in Konigsberg/Zürich discovered the capacitive properties of tissue and the anisotropy of muscle conductance based on alternating current measurements. In 1864, James C. Maxwell contrived his now famous equations by specifically calculating the resistance of a homogeneous suspension of uniform spheres as a function of their volume concentration (1). In 1928, Kenneth S. Cole expanded on Maxwell's model by determining

the impedance of a suspension of capacitively coated spheres over a range of frequencies.

Several additional and important developments occurred before the start of World War II. Rudolph Hoebers studied the conductivity of blood and found it to be dependent on the stimulation frequency. Simultaneously, the electrical properties of proteins and amino acids were discovered and extensively studied by Oncley, Fricke, and Wyman (3). These contributions lead to further developments in the science of biophysics and electrophysiology.

Bioimpedance research accelerated after World War II. In 1950, Nyboer et al. launched an investigation into thoracic electrical bioimpedance (TEB) as an alternative to invasive methods of measuring cardiac function and published a novel method termed "Impedance Plethysmography" (4,5). However, Kubicek and Patterson were credited with the development of the original TEB system in conjunction with the National Aeronautics and Space Administration in the mid-1960s (6). This device was designed to monitor stroke volume (SV) and cardiac output (CO) noninvasively during space flight. In addition, Djordjevich and Sadove coined the term "electrohemodynamics" in 1981 to describe a science that relates the theories of fluid mechanics and elasticity to the continuous impedance signal and to the time variations of arterial blood pressure (7). Jan Baan et al. introduced the impedance or conductance catheter technique to measure real time chamber volume in the mid-1980s (8). This technique revolutionized the study of cardiovascular mechanics in both the laboratory and clinical settings by making the study of ventricular-pressure volume relationships practical.

More recently, bioimpedance applications have continued to expand, especially in the area of implantable devices. Modern pacemakers and defibrillators routinely use biompedance measurements to verify pacing lead performance and position, monitor minute ventilation and thoracic fluid content, and optimize programmable device features such as pacing rate and AV delay in a closed-loop fashion (9–11).

The terms bioimpedance or tissue impedance describe both the resistive and reactive components of tissue at the applied stimulus frequency. The capacitive reactive components of the measured tissue impedance change at higher frequencies due to the relative conductive properties of tissue fluids and cellular membranes. Bioimpedance methods can be categorized into two areas: impedance plethysmography and impedance cardiography. Impedance plethysmography, by definition, refers to the measurement of a volume change in a heterogeneous tissue segment using electrical impedance in which the changing impedance waveform (ΔZ) is used to determine cardiac, respiration, and peripheral volume change as a function of time. In contrast, impedance cardiography is a subdivison of impedance plethysmography that focuses on the measurement of cardiac stroke volume and widely uses the first derivative (dZ/dt) of the changing impedance waveform (ΔZ) to monitor fiducial time element points such as cardiac valve opening and closing. Both methods primarily use a single low frequency stimulus current (<100 kHz) where most of the elements in the current paths are primarily resistive. Techniques such as impedance plethysmography and impedance cardiography primarily depend on resistive rather than reactive components of the blood impedance. Thus, applications using low frequency stimulus current to primarily measure the resistive component of bioimpedance will be categorized in this article as resistive applications of bioimpedance.

The second general category of bioimpedance measurement involves estimation of fluid volume distributions such as intracellular and extracellular volume, percent body fat vs. percent muscle mass, and cell and tissue viability. This area primarily employs a multifrequency stimulus current bandwidth (>1000 Hz) where most of the elements in the current path contain significant resistive and reactive components. Thus, applications using high frequency stimulus current to measure the resistive and reactive components of bioimpedance will be categorized as reactive applications of bioimpedance.

Bioimpedance Theory

When constant electric current is applied between two electrodes through a biological medium and the corresponding voltage is measured between the two source poles, the resultant impedance or bioimpedance is determined by Ohm's law. The recorded voltage is the sum of the potential difference contributions due to the electrical conductivity properties of the tissue medium. The exchange of electrons from source to sink occurs from electrons of the metal electrode (such as platinum or silver-silver chloride) to ions of the tissue medium. The electrode is the site of charge carrier exchange between electrons and ions and thus serves as a transducer of electrical energy. Impedance measurements most commonly use a two-electrode (bipolar) or four-electrode (tetrapolar) arrangement (Fig. 1). In the bipolar arrangement, the electrodes serve as both the current source (anode and cathode, respectively) and as the measurement electrodes. In the tetrapolar arrangement, one electrode serves as current source anode, one as the current source cathode, and the remaining two electrodes serve as the respective measurement electrodes. A disadvantage of the bipolar electrode system is electrode polarization due to a frequency-dependent polarization impedance. Therefore,

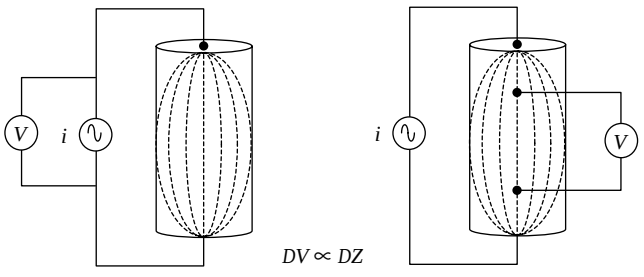


Figure 1. Bipolar (left) and tetrapolar (right) electrode configurations. With a constant current stimulation source, the change in measured voltage (ΔV) is proportional to the change in calculated impedance (ΔZ). Tetrapolar systems require additional electrodes but avoid electrode polarization effects.

the measured voltage in a bipolar impedance system reflects the combined impedance of both the tissue segment and the electrode tissue interface. When the voltage is measured with an isolated high input impedance electrode system, such as with the tetrapolar lead configuration, minimal current flows in the isolated sensing electrodes, thus problems with electrode polarization can be effectively reduced (12).

Endogenic ionic current movement between and within cellular structures encompasses the electrical properties of tissues and the term bioelectricity. In tissue and the living cell, an inseparable alliance exists between electricity and chemistry (1). The perception of current through human tissue is dependent on frequency, current density, effective electrode area, and current duration. The maximum sensitivity of the nervous system is approximately in the range of 10 to 1000 Hz for sine waves. At frequencies greater than 1 kHz, the sensitivity is strongly reduced.

Measured bioimpedance is a function of the real and reactive components of the tissue medium at the applied frequency of the stimulus current. Tissue characteristics at low frequencies are almost independent of cellular membrane reactance and internal intracellular resistivity. Thus, most of the applied current is conducted via the extracellular fluid. The cellular membrane behavior at intermediate or high frequencies is primarily a characteristic of membrane reactance and internal resistivity. The membranes are an impure reactance and, therefore, show a dielectric loss and a phase angle, which is independent of frequency (13). At higher frequencies, the cell's membrane reactance and resistance become negligible and the applied current is conducted through both the intracellular and extracellular fluid.

RESISTIVE APPLICATIONS OF BIOIMPEDANCE

Transthoracic Bioimpedance

The Cylindrical Model. Many applications of bioimpedance measurement focus on primarily resistive changes. These techniques are all based on the cylindrical model (Fig. 2) represented by a tissue volume with uniform cross-sectional area (A), length (L), and resistivity (ρ) (14).

$$R = \rho \frac{L}{A} \tag{1}$$

The resistance of the vessel segment is directly proportional to the resistivity of the conductive medium and length of the vessel segment and inversely proportional to the cross-sectional area of the vessel segment (Eq. 1). As shown in Fig. 2, as the cross-sectional area of the vessel segment increases from A_1 to A_3, the measured resistance decreases. Resistivity (ρ) is a tissue property that varies substantially between tissues. Typical tissue resistivities are shown in Table 1 (15).

Equation 1 can be modified to determine the volume of the tissue by multiplying both sides of the equation by L, substituting impedance (Z) for resistance, and solving for volume (V) as a function of time (Eq. 2):

$$V(t) = \rho \frac{L^2}{Z(t)} \tag{2}$$

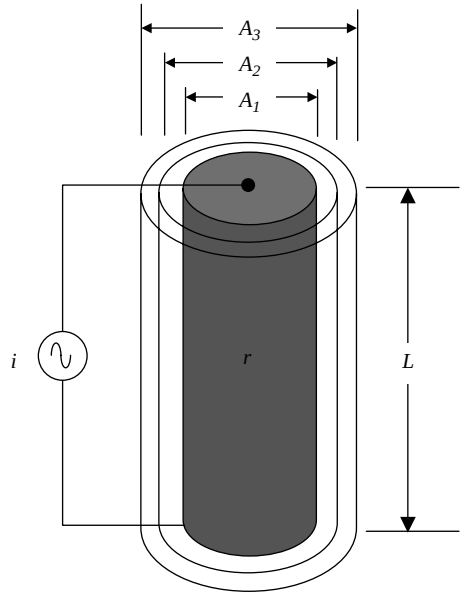


Figure 2. Cylindrical model of a vessel segment. A_1–A_3 represent cross-sectional area changes of the tissue of the interest (e.g., blood). Tissue length (L) is often determined by measurement electrode spacing. Blood resistivity (ρ) also determines the resistance (R) of the tissue volume. Resistance measured is directly proportional to the measured voltage and indirectly proportional to the constant alternating current (i) applied to the vessel segment.

The cylindrical model is based on several important assumptions: The electrical field, and hence current density, is homogeneous within the tissue of interest, the current is completely confined to the tissue of interest, and the values shown in Table 1 do not account for tissue anisotropy. Most biological tissues have lower resistivity in the longitudinal direction of cell or fiber orientation (12,16–18). For example, the ratio of resistivity in the transverse to parallel direction can be > 3 in cardiac tissue (18). These assumptions may not be valid for some applications of bioimpedance such as the conductance catheter technique for chamber volume estimation (see below). Therefore, the

Table 1. Various Tissue Resistivities in ohms-meter ($\Omega \cdot$ m) (15)

Tissue	ρ ($\Omega \cdot$ m)
Blood (Hematocrit = 45)	1.6
Plasma	0.7
Heart Muscle (Longitudinal)	2.5
Heart Muscle (Transverse)	5.6
Skeletal Muscle (Longitudinal)	1.9
Skeletal Muscle (Transverse)	13.2
Lung	21.7
Fat	25

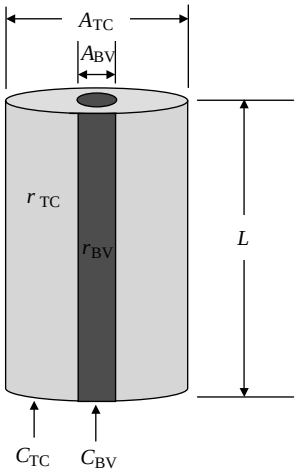


Figure 3. Parallel-column model of the thoracic cavity. This two-column model (C_{TC} and C_{BV}) represents a thoracic cavity segment of length (L), cross-sectional areas of the great blood vessels (A_{BV}), and thoracic cavity segment (A_{TC}), and resistivities of the thoracic cavity segment tissues (ρ_{TC}) and blood volume (ρ_{BV}).

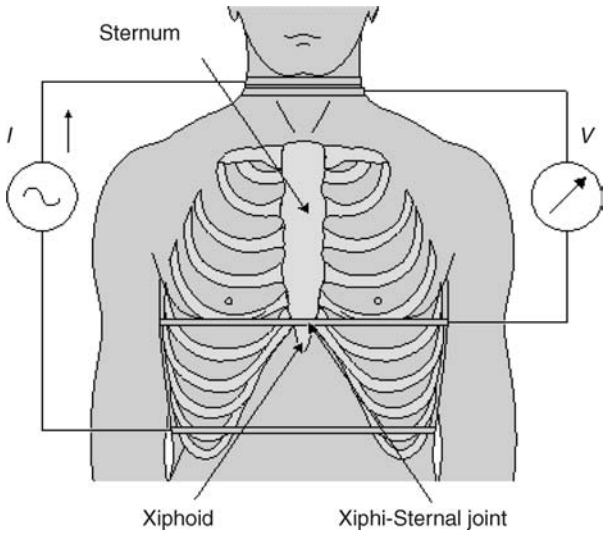


Figure 4. Transthoracic band electrode placement for stroke volume estimates. The two outer-band electrodes supply the stimulus current (I); the two inner electrodes measure the corresponding voltage (V). Impedance is calculated from the ratio of V/I (15).

cylindrical model must be adjusted for particular applications.

The Parallel-Column Model. The parallel-column model, first described by Nyboer (5) (Fig. 3) is closely related to the cylindrical model, but accounts for current leakage into surrounding tissues. The model consists of a smaller cylindrical conductor (C_{BV}) of length L representing the large blood vessels of the thoracic cavity (i.e., aortic and pulmonary arteries) embedded in a larger cylindrical conductor (C_{TC}) of the same length (L) representing the tissues of the thoracic cavity. C_{BV} consists of blood with specific resistivity (ρ_{BV}) and time-varying cross-sectional area (A_{BV}). C_{TC} is assumed to be heterogeneous (i.e., bone, fat, muscle) with specific resistivity (ρ_{TC}) and constant cross-sectional area (A_{TC}). Thus, the cylindrical model of the time-varying volume can be modified:

$$V_T(t) = \rho_{RV}\frac{L^2}{Z_{BV}(t)} + \rho_{TC}\frac{L^2}{Z_{TC}} \tag{3}$$

where $V_T(t)$ represents the total volume change. As the distribution of the measured resistance and the net resistivity of the parallel tissues are unknown, calculation of absolute volume can be problematic. However, the constant volume term drops out when the change in volume is calculated from Eq. 3:

$$\Delta V_{BV}(t) \cong \rho_{BV}\left(\frac{L^2}{Z_0^2}\right)\bullet\Delta Z(t) \tag{4}$$

where Z_0 is the basal impedance measured and $\Delta Z(t)$ is the pulsatile thoracic impedance change. Thus, Eq. 4 links the parallel cylindrical model to TEB estimates of stroke volume (ΔV_{BV}).

Noninvasive Measurement of Cardiac Output. Transthoracic electrical bioimpedance (TEB) was first introduced by Patterson et al. in 1964 (19). As shown in Fig. 4 (13), this system employs two pairs of band electrodes positioned at the superior and inferior ends of the thorax in the cervical and substernal regions, respectively. The outer electrode pair drives a constant current (I) and the inner electrode pair is used to measure the corresponding voltage (V), which is a function of the varying impedance changes during respiration and the cardiac cycle.

Noninvasive measurements of stroke volume can be determined with the configuration in Fig. 4 by applying Eq. 5.

$$SV = \left[\rho_{BV}\left(\frac{L^2}{Z_0^2}\right)\bullet\Delta Z(t)\right]\bullet LV_{ET} \tag{5}$$

Cardiac output may then be determined by multiplying stroke volume (SV) by heart rate (HR). $\Delta Z(t)$ is the measured time-varying impedance signal, and Z_0 represents the nonpulsatile basal impedance.

Sramek et al. (20) modified Patterson et al.'s (19) parallel cylinder model into a truncated cone in order to improve stroke volume predictions (Eq. 6). The physical volume of the truncated cone was determined to be one-third the volume of the larger thoracic cylinder model.

$$SV = \left(\frac{(L)^3}{4.2}\right)\bullet LV_{ET}\bullet\left(\frac{(dZ/dt)_{max}}{Z_0}\right) \tag{6}$$

Sramek et al. (20) also found that in a large normal adult population, the measured linear distance (L) is equal to 17% of body height (cm). Cardiac output is directly proportional to body weight (21). As ideal body weight is a linear function of overall height (22), the proportionality of height (H) to cardiac output can be represented in the first term of Eq. 6 by $(0.17H)^3/4.2$.

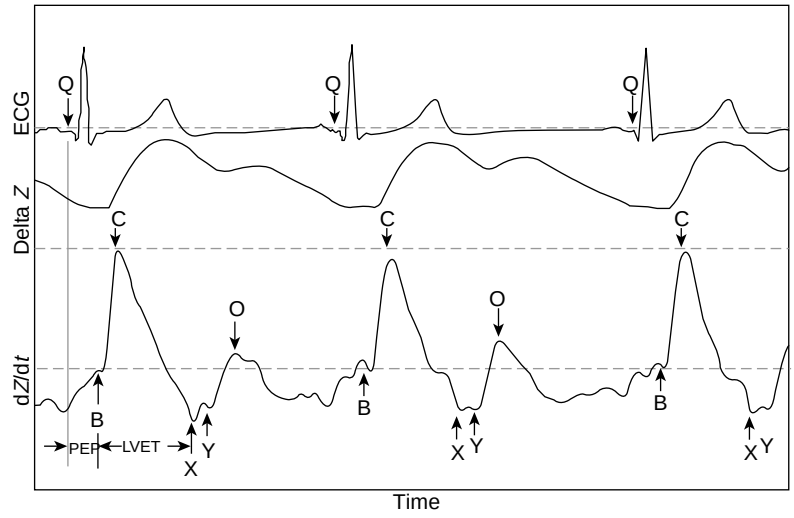


Figure 5. Impedance cardiography waveforms. Three waveforms depict the electrocardiogram (ECG), transthoracic impedance change as a function of time (ΔZ), and first-time derivative of the impedance change dZ/dt. Impedance waveforms are intentionally inverted to show a positive deflection during cardiac contraction. Fiducial points on the dZ/dt waveform are represented by the opening of the aortic and pulmonic valves (B), closure of the aortic (X) and pulmonic (Y) valves, mitral valve opening (O), ventricular pre-ejection period (PEP), and left ventricular ejection time (LVET). Q represents the end of atrial contraction (46).

Other empirical modifications to the original model have also been proposed in order to improve CO and SV estimates (20,22–28). In addition, various modifications to the external lead configuration have also been proposed to improve SV estimates, including the application of tranesophageal electrodes (29,30).

Several commercial bioimpedance systems are available for clinical noninvasive estimation of cardiac output and other hemodynamic parameters. The advantages of such systems include noninvasive application, relatively low cost, and lack of noninvasive alternatives. However, these techniques have gained somewhat limited clinical acceptance due to suspect reliability over a wide range of clinical conditions.

A myriad of validation studies of TEB estimates of cardiac output have been published with equivocal results (6,7,19,24,28,31–40). For example, Engoren et al. (35) recently compared cardiac output as determined by bioimpedance, thermodilution, and the Fick method and showed that the three methods were not interchangeable in a heterogenous population of critically ill patients. Their data showed that measurements of cardiac output by thermodilution were significantly greater than by bioimpedance. However, the bioimpedance estimates varied less than the thermodilution estimates for each subject. In contrast, a meta-analysis of impedance cardiography validation trials by Raaijmakers et al. (38) showed an overall correlation between cardiac output measurements using transthoracic electrical bioimpedance cardiography and a reference method of 0.82 (95% CI: 0.80–0.84). The performance of impedance measurement of cardiac output was similar in various groups of patients with different diseases with the exception of cardiac patients, in which group the correlation was decreased. Additional investigations by Kim et al. (41) and Wang et al. (42) used a detailed 3D finite element model of the human thorax to determine the origin of the transthoracic bioimpedance signal. Contrary to the theory that lead to the parallel column model formulae, these investigators determined that the measured impedance signal was determined by multiple tissues and other factors that make reliable estimates of cardiac output over a wide variety of physiological conditions difficult. Nevertheless, commercially available impedance plethysmographs provide estimates of cardiac output that may be useful for assessing relative changes in cardiac function during acute interventions, such as optimization of implantable pacemaker programmable options such as AV delay (11).

Cardiac Cycle Event Detection. TEB also focuses on measurements of the change in impedance (ΔZ) and the impedance first time derivative (dZ/dt) measured simultaneously with the electrocardiogram (ECG). Figure 5 depicts a typical waveform of the aforementioned parameters. Note that the impedance change (ΔZ) and the impedance first time derivative (dZ/dt) are inverted by convention (43). The value of dZ/dt is measured from zero to the most negative point on the waveform. The ejection time (LV_{ET}) is an important parameter in determining stroke volume (Eqs. 5 and 6). This systolic time interval allows an estimation of cardiac contractility. The Heather Index (HI) is another proposed index of contractility from systolic time intervals determined by the dZ/dt waveform (33,44) (Eq. 7):

$$HI = \frac{dZ/dt_{\max}}{QZ_1}$$

where $dZ/dt_{\max}$ (point C) is the maximum deflection of the initial waveform derived from the ΔZ waveform and QZ_1 is the time from the beginning of the Q wave to peak $dZ/dt_{\max}$ (Fig. 5). In this figure, point Q represents the time between the end of the ECG p-wave (atrial contraction) and the beginning of the QRS wave (ventricular depolarization) (45). Point B depicts the opening of the aortic and pulmonic valves. After the ventricles depolarize and eject the blood volume into the aortic and pulmonary arteries, points X and Y represent the end systolic component of the cardiac cycle as closure of the aortic and pulmonic valves, respectively.

Passive mitral valve opening and passive ventricular filling begins at point O. Although the timing of the various

fiducial notches of the dZ/dt waveform is well known, controversy remains with the origins of the main deflections and are not well understood (15).

Signal Noise. In TEB measurements, several filtering techniques have been proposed to attenuate undesired noise sources depending on which component of the impedance waveform is desired (i.e., respiratory, cardiac, or mean impedance) (47). Most of the signal processing techniques for impedance waves use ensemble averaging for the elimination of motion artifacts (48). A recent signal processing technique described by Wang et al. (48) uses the time-frequency distribution to identify fiducial points on the dZ/dt signal for the computation of left ventricular ejection time and dZ/dt_{max}. As shown in Fig. 5, many of the fiducial points on the dZ/dt waveform are clearly identifiable, but may be somewhat more difficult to observe under severe interference conditions.

Filtering techniques have also been proposed to eliminate noise caused by respiration such as narrow band-pass filtering around the cardiogenic frequency. However, such filtering techniques often eliminate the high frequency components of the cardiac signal and introduce phase distortion (49). To help alleviate this problem, various techniques to identify breathing artifacts with forward and backward filtering have been employed (50,51). Despite these techniques, motion artifact remains with unknown frequency spectra that may overlap the desired impedance frequency spectra during data acquisition. Adaptive filters represent another approach and may eliminate the motion artifact by tracking the dynamic variations and reduce noise uncorrelated to the desired impedance signal (49). Raza et al. (51) developed a method to filter respiration and low frequency movement artifacts from the cardiogenic electrical impedance signal. Based on this technique, the best range for the cutoff frequency appears to be from 30–50% of the heart rate under supine, sitting, and moderate exercise conditions (51).

Applications of Transthoracic Bioimpedance. ***Hypertension.*** TEB has emerged as a noninvasive tool to assess hemodynamic parameters, especially within the framework of hypertension monitoring (24–28,52–54). Measurement of the various hemodynamic components such as stroke volume, ejection time, systemic vascular resistance, aortic blood velocity, thoracic fluid content, and contractility (i.e., Heather Index) using impedance cardiography in patients with hypertension allows more complete characterization of the condition, a greater ability to identify those at highest risk, and allows more effectively targeted drug management (25,53). Several studies have used TEB to evaluate hemodynamic parameters and demonstrated that TEB-guided therapy improves blood pressure control (53,55,56). For example, in a three month clinical study by Taler et al. (55) 104 hypertensive patients were randomized to either TEB-guided therapy or standard therapy. The results showed improved blood pressure control in the TEB-guided group. The investigators concluded that measurement of hemodynamic parameters with TEB methods was more effective than clinical judgment alone in guiding selection of antihypertensive therapies in patients resistant to empiric therapy (53).

Pacemaker Programming. TEB estimates of cardiac index technique has been investigated as a noninvasive method to optimize AV delay intervals in pacemaker patients in an open-loop fashion (57,58). Ovsyshcher et al. (58) measured stroke volume changes at various programmed AV delays via impedance cardiography in dual-chamber pacemaker patients. The optimal and worst programmed AV delays were identified as the settings that produced the highest and lowest cardiac index, respectively. As shown in Fig. 6 (58), the highest cardiac index values resulted with mean AV delays <200 ms and the lowest cardiac index values resulted with mean AV delays >200 ms.

More recently, a study by Tse et al. (59) evaluated AV delay interval optimization during permanent left ventricular pacing using transthoracic impedance cardiography in conjuction with Doppler echocardiography over a range of AV intervals. This study revealed no significant difference between the optimal mean AV delay interval determined by transthoracic impedance cardiography and that determined by Doppler echocardiography. However, as shown in Fig. 7, the mean cardiac output at different AV

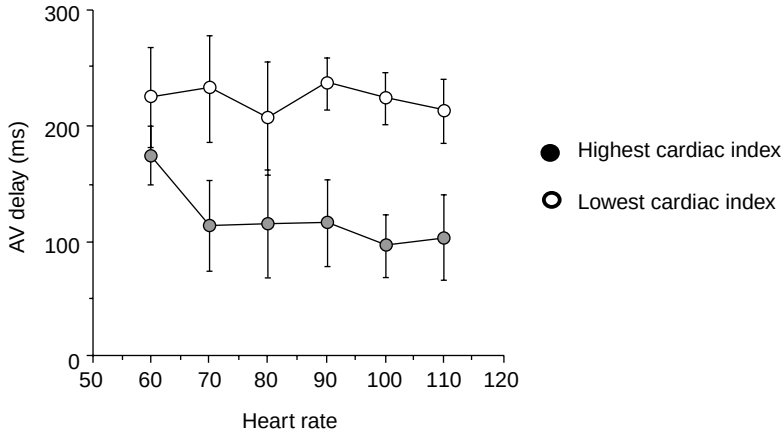


Figure 6. Highest and lowest cardiac indices at varied AV delays and pacing rates. Highest cardiac index values (closed circles) resulted with mean AV delays <200 ms and the lowest cardiac index values (open circles) resulted with mean AV delays >200 ms (58).

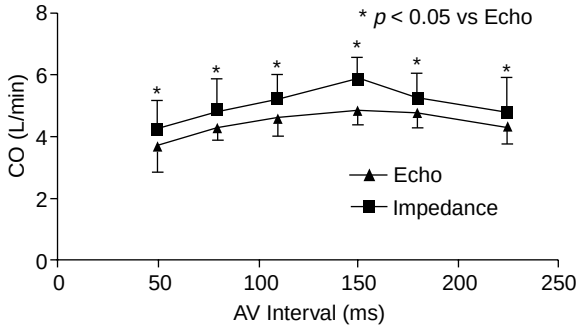


Figure 7. Cardiac output measured by impedance cardiography and Doppler echocardiography at various AV delay intervals (59).

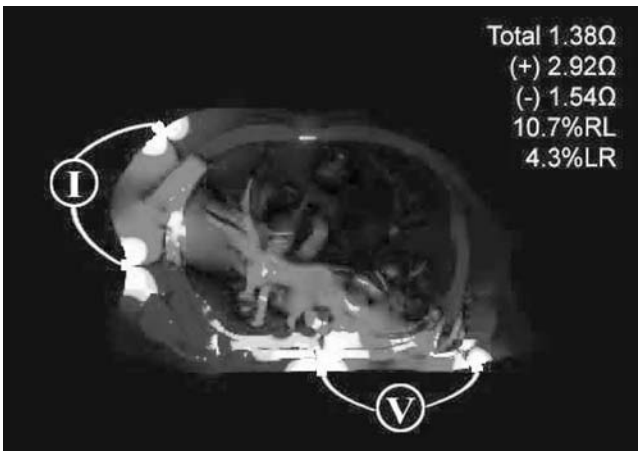


Figure 9. High resolution computer simulated model of the thorax. Regions of both positive and negative sensitivity contribute to the total impedance measured. Negative impedance sensitivity regions, (1.54 Ω) and positive impedance sensitivity regions (red, 2.92 Ω) both contribute to the measure total impedance of 1.38 Ω. RL = contribution of right lung to measured impedance. LR = concontribution of left lung to measured impedance (62).

delay intervals was significantly higher when measured by transthoracic impedance cardiography than when measured by Doppler echocardiography.

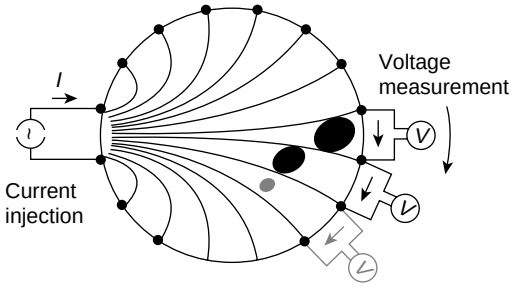
Electrical Impedance Tomography. Electrical impedance tomography (EIT) is a technique to reconstruct low resolution cross-sectional images of the body based on differential tissue resistivity (60). The image is created using an array of 16–32 electrodes, usually positioned around the thorax (Fig. 8). Impedance is computed from all electrodes as the drive electrodes rotate sequentially about the tissue surface. The "image" is then reconstructed using standard tomographic techniques. The advantages of EIT include low cost and the potential for ambulatory applications. The disadvantages include the low resolution of the image, the contribution of "out-of-plane" tissues to the "in-plane" image, and the limited clinical applications.

Recent improvements in hardware and software systems that increase the accuracy and speed of regional lung volume change have maintained interest in this technology (60–62). Besides pulmomary monitoring, other potential applications of EIT include neurophysiology, stroke detection, breast cancer detection, gastric emptying, and cryosurgery (63–66).

Lead Field Theory. An analysis of sensitivity is crucial to interpretation and application of EIT images as well as other bioimpedance applications. The sensitivity distribution of an impedance measurement provides the relation between the measured impedance resulting from the conductivity distribution of the measured region. It describes the relative contribution of each region to the measured impedance signal. The contribution of any region to the measurement is not always intuitively obvious and the magnitude of the sensitivity may be less than zero (Fig. 9). Therefore, the relative contribution of various tissues to the reconstructed "image" can be difficult to interpret.

The applicability of lead field theory in impedance measurements has been shown theoretically by Geselowitz (67). According to that theory, appropriate selection of the electrode configuration enables increased measurement sensitivity and selectivity to particular regions (68). Also, the measured impedance change (ΔZ) can be evaluated from the change in conductivity within a volume conductor $\Delta\sigma$ and the sensitivity distribution S by (67):

$$\Delta Z = \int_v \frac{1}{(\Delta)\sigma} \bullet S_{dv} \tag{8}$$



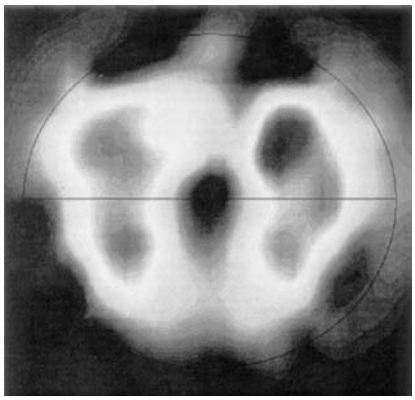


Figure 8. Left: Cartoon representation of a system to generate an electrical impedance tomographic image: 16 electrodes around the chest inject currents and record the resultant voltage in a sequential manner. Right: Electrical impedance tomographic image of the thoracic cavity. Heart and lung tissue are distinguishable (60).

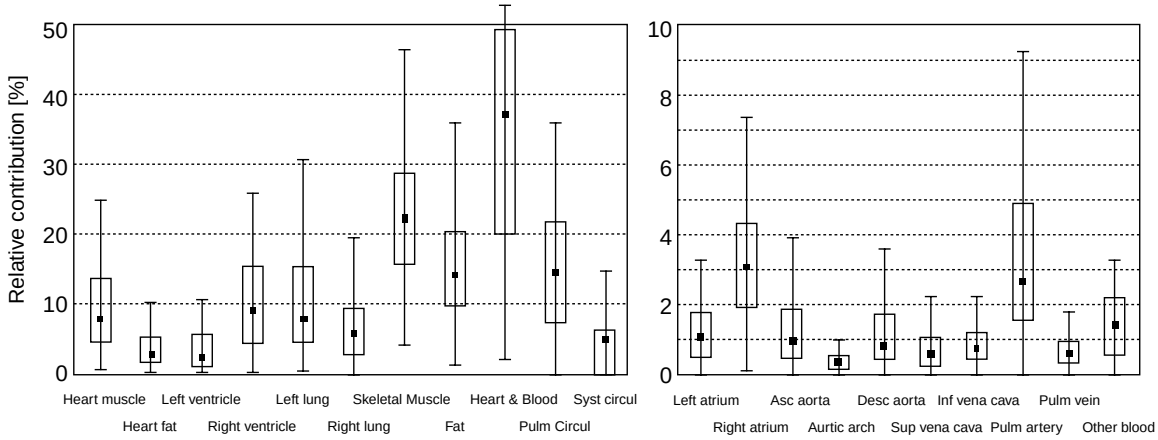


Figure 10. Simulated measurement sensitivities of tissues. Values are indicated for each tissue type in addition to three tissue groups consisting of pulmonary circulation, systemic circulation, and all the blood masses and heart muscle (68).

The measurement sensitivity S is obtained by first determining the current fields generated by a unit current applied to the current injection electrodes and the voltage measurement electrodes. These two lead fields form the combined sensitivity field of the impedance measurement associated with the electrode configuration by:

$$S = J_{\mathrm{LE}} \bullet J_{\mathrm{LI}} \tag{9}$$

where:

S = the scalar field giving the sensitivity to conductivity changes at each location,

J_{LI} = the lead field produced by current excitation electrodes,

J_{LE} = the lead field produced by voltage measurement leads.

Therefore, sensitivity at each location depends on the angle and magnitude of the two fields and can be positive, negative, or null. The relative magnitude of the sensitivity field in a tissue segment provides a measure of how conductivity variation in that tissue segment will affect the detected ΔZ (69).

Lead field theory suggests that the relative contribution of a tissue to the measured impedance depends on the properties of the tissue, the symmetric arrangement of the tissues, and the geometry of the applied current and voltage electrodes. The precise relative contribution of various tissues to measure impedance is therefore difficult to predict (Fig. 10) (68).

Intrathoracic Bioimpedance

Minute Ventilation. As described earlier, respiratory rate can be estimated with TEB. However, intrathoracic impedance sensing has also been applied to measure respiratory rate and minute ventilation in implantable devices such as pacemakers and implantable cardiac defibrillators (ICDs). Intrathoracic impedance vector configurations typically consist of a tripolar arrangement with bipolar pacing or ICD leads placed in the right ventricle (RV) and the device "can" (metal case or housing) placed subcutaneously in the left or right pectoral region. A variety of anode/cathode electrode arrangements are possible with current source electrodes such as the proximal electrode (RV-ring) to can or the distal electrode (RV-tip) to can and voltage sense electrodes between RV-coil to can. Typically, a low energy pulse of low current amplitude (1 mA with pulse duration of 15 μs) is delivered every 50 milliseconds (10). Figure 11 depicts a typical lead arrangement used for intrathoracic impedance measurements.

The electric fields generated with this electrode configuration must be arranged to intersect in parallel in order to provide the greatest sensitivity. The sensitivity of an electrode is proportional to the current density of the applied stimulus. Moreover, the sensitivity is highest close to the current-injecting electrodes and lowest toward the center of the tissue medium within the lead field vector,

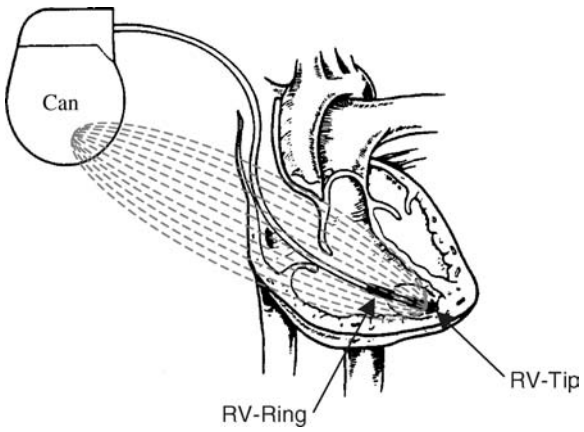


Figure 11. Lead configuration for intracardiac impedance measurments. Stimulus current injected from RV-tip to can. Voltage is measured from RV-Ring to can. The large size of the can reduces electrode polarization effects of the tripolar lead configuration.

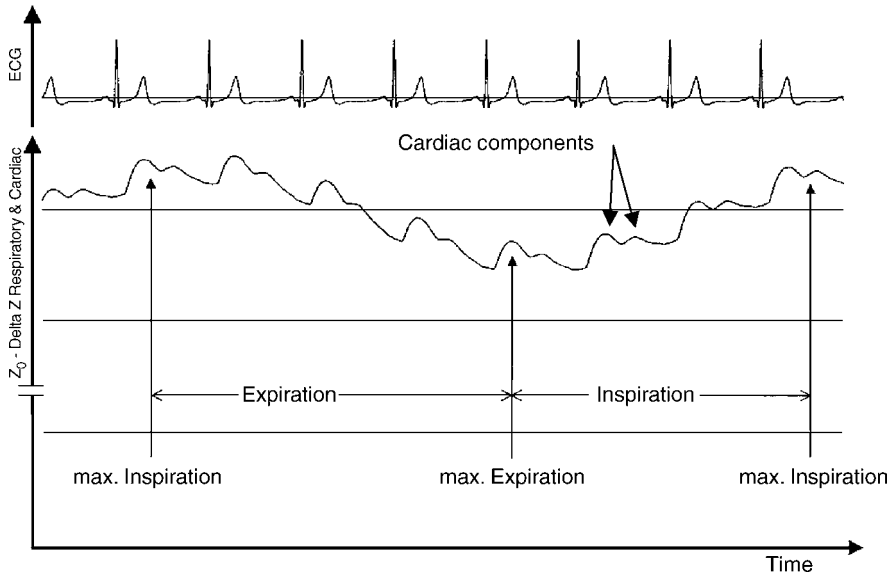


Figure 12. Respiratory variation in impedance waveform. Electrocardiogram (ECG) shown with ΔZ waveform. ΔZ waveform is comprised of higher frequency cardiac components superimposed on the lower frequency respiratory variation component (46).

because the applied field current density is lowest in this region.

Experimental evidence indicates that the frequency and amplitude of the respiratory component of the bioimpedance signal are related to changes in both the respiratory rate and the tidal volume and, hence, the minute ventilation (MV). MV sensing in rate-adaptive pacing systems has also been shown to closely correlate with carbon dioxide production ($V\mathrm{CO}_2$) (10). This relationship has been applied in some commercially available pacemakers with automatic rate-adaptive pacing features (9,10).

As shown in Fig. 12 (46), the amplitude of impedance changes during respiration are significantly larger than the higher frequency cardiac components. By magnitude, the change in the cardiac component of the impedance waveform is in the range of 0.1–0.2 Ω, which correlates to approximately 0.3–0.5% of the thoracic impedance (ΔZ) (46). Moreover, each component has a different frequency, typically 1.0–3.0 Hz for cardiac activity and 0.1–1.0 Hz for respiratory activity (9). This differentiation allows extraction of each signal by specific filtering techniques.

In general, the minute-ventilation sensor is characterized by a highly proportional relationship to metabolic demand over a wide variety of exercise types (10). However, optimal performance of impedance-based MV sensors to control pacing rate during exercise often requires careful patient-specific programming.

Fluid Status. Fluid congestion in the pulmonary circulation due to volume overload results in preferential transport of fluid primarily into the extracelluar fluid space and not into the intracellular compartments. Clinical symptoms to assess fluid overload include hypertension, increased weight, pulmonary or peripheral edema, dyspnea, and left ventricular dysfunction. Recently, implantable device-based bioimpedance measurements have been applied to detect thoracic fluid accumulation in patients with congestive heart failure (CHF) and to provide early warning of decompensation caused by factors such as volume overload and pulmonary congestion (32,70,71). This application is the result of a substantial body of new and historical experimental evidence (32,70–73).

Externally measured transthoracic impedance techniques have been shown to reflect alterations in intrathoracic fluid and pulmonary edema in acute animal and human studies (72). The electrical conductivity and the value for transthoracic impedance are determined at any point in time by relative amounts of air and fluid within the thoracic cavity (73). Additional studies have suggested that transthoracic impedance techniques provide an index of the fluid volume in the thorax (32,71). Wang et al. (70) employed a pacing-induced heart failure model to demonstrate that measurement of chronic impedance using an implantable device effectively revealed changes in left ventricular end-diastolic pressure in dogs with pacing-induced cardiomyopathy (Fig. 13) (70). Several factors were identified that may influence intrathoracic impedance with an implantable system, including (*1*) fluid accumulation in the lungs due to pulmonary vascular congestion, pulmonary interstitial congestion, and pulmonary edema; (*2*) as heart failure worsens, heart chamber dilation and venous congestion occur and pleural effusion may develop; and (*3*) after implant, the tissues near the pacemaker pocket swell and surgical trauma can cause fluid buildup (70).

Yu et al. (74) also showed that sudden changes in thoracic impedance predicted eminent hospitalization in 33 patients with severe congestive heart failure (NYHA Class III–IV). During a mean follow-up of 20.7 ± 8.4 months, 10 patients had a total of 25 hospitalizations for worsening heart failure. Measured impedance gradually decreased before admission by an average of $12.3 \pm 5.3\%$ ($p < 0.001$) over a mean duration of 18.3 ± 10.1 days. The decline in impedance also preceded the symptom onset by a mean lead time of 15.3 ± 10.6 days ($p < 0.001$). During hospitalization, impedance was inversely correlated with

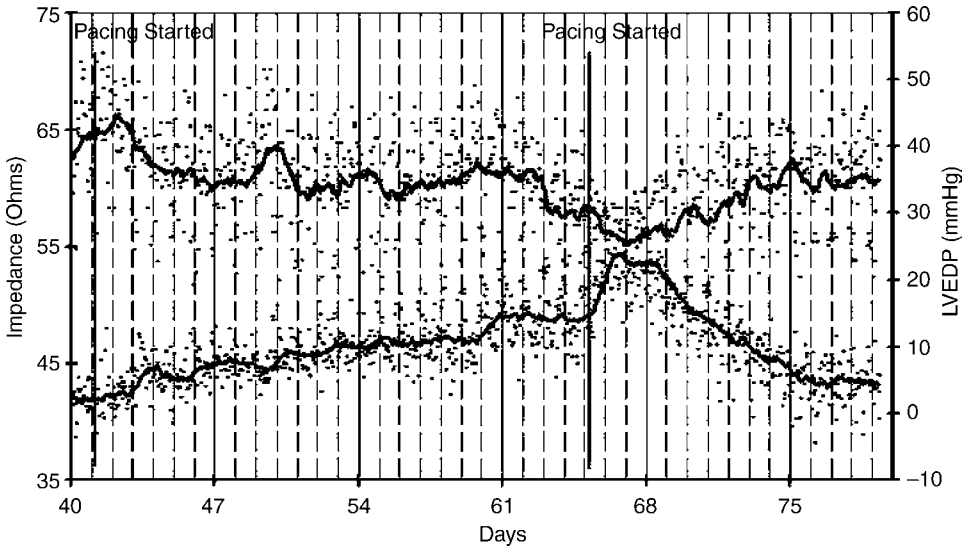


Figure 13. Impedance vs. LVEDP during pacing-induced heart failure. Intrathorcic impedance via an implantable device-lead configuration and LVEDP are inversely correlated in a canine model of pacing-induced cardiomyopathy. A general trend for impedance to decrease as heart failure developed is shown. Once pacing induced heart failure was terminated, LVEDP and impedance returned to basal levels (70).

pulmonary wedge pressure (PWP) and volume status with $r = -0.61$ ($p < 0.001$) and $r = -0.70$ ($p < 0.001$), respectively. Automated detection of impedance decreases was 76.9% sensitive in detecting hospitalization for fluid overload with 1.5 false-positive (threshold crossing without hospitalization) detections per patient-year of followup. Thus, intrathoracic impedance from the implanted device correlated well with PWP and fluid status, and may predict eminent hospitalization with a high sensitivity and low false-alarm rate in patients with severe heart failure (Fig. 14) (74). Some commercially available implantable devices for the treatment of CHF or ventricular tachyarrhythmias now continually monitor intrathoracic impedance and display fluid status trends. This information is then provided to the clinician via direct-device interrogation or by remote telemetry.

Volume Conductance Catheter. The conductance catheter technique, first described by Baan et al. (8), enables continuous measurements of chamber volume, particularly left ventricular (LV) volume. This method has been used extensively to assess global systolic and diastolic ventricular function (75). In many respects, the conductance catheter revolutionized the study of cardiovascular mechanics in both the laboratory and clinical settings by making the study of ventricular pressure-volume relationships practical. The technique led to a renaissance of cardiac physiology over the past 25 years (76) by increasing the understanding of the effect of pharmacologic agents, disease states, pacing therapies, and other interventions on cardiovascular function. Conductance catheter systems are available for clinical and laboratory monitoring applications, including a miniature system capable of measuring LV volume and pressure in mice (77).

The conductance methodology is based on the parallel cylinder model (Fig. 3). However, the cylindrical model assumes that the volume of interest has a uniform cross-sectional area across its length. Therefore, the ventricular volume is subdivided into multiple segments determined by equipotential surfaces bounded by multiple sensing electrodes along the axis of the conductance catheter (Fig. 15).

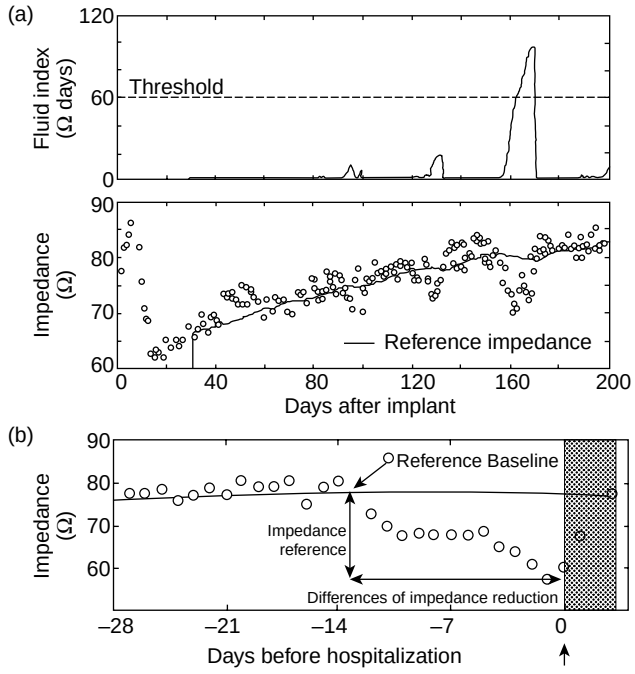


Figure 14. Fluid status monitoring with an implanted device. A: Operation of algorithm for detecting decreases in impedance over time. Differences between measured impedance (bottom; °) and reference impedance (solid line) are accumulated over time to produce fluid index (top). Threshold values are applied to fluid index to detect sustained decreases in impedance, which may be indicative of acutely worsening thoracic congestion. B: Example of impedance reduction before heart failure hospitalization (arrow) for fluid overload and impedance increase during intensive diuresis during hospitalization. Label indicates reference baseline (initial reference impedance value when daily impedance value consistently falls below reference impedance line before hospital admission). Magnitude and duration of impedance reduction are also shown. Days in hospital are shaded (74).

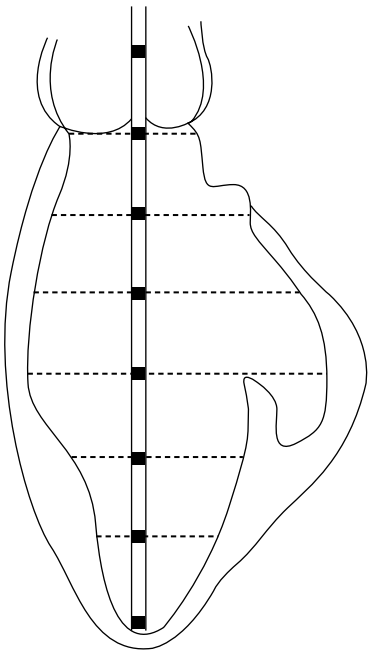


Figure 15. Conductance catheter modeled in the left ventricle (LV). Stimulus current is injected from the proximal and distal catheter electrodes. Voltage is measured between the remaining adjacent electrode pairs. Total conductance is calculated by the summation of all segmental conductances measured in the individual segments.

The two most distal electrodes are used to generate an electric field, typically 0.4 mA p-p, at 20 kHz. The remaining electrodes are used in pairs to measure the conductance of several segments ($n =$ number of segments), which represent the instantaneous volumes of the corresponding segment. The conductance is then converted to volume by modifying Eq. 2:

$$V(t) = \rho L^2 \sum_{i=1}^{n} G_i(t) \tag{10}$$

where G is the time-varying conductance of segment i. However, the conductance technique also violates two other key assumptions of the cylindrical model. First, the electrical field generated by the drive current electrodes is not homogenous and, second, the electric field is not confined to the chamber of interest (i.e., the LV). Thus, the multiple segment cylindrical model has been modified in order to allow conductance catheter estimates of volume to agree with gold standard estimates such as echocardiography (Eq. 11):

$$V(t) = \left(\frac{\rho L^2 \Sigma G(t)}{\alpha}\right) - V_P \tag{11}$$

where correction factor α accounts for nonhomogeneity of the electric field and the correction factor V_P accounts for the current leakage into the surrounding tissues. The terms α and V_p are related and may vary somewhat during the cardiac cycle (16,78). Various methods have been applied to determine the values of α and V_P, including the method of hypertonic saline injection. Recently, the concept of dual-frequency excitation has been applied to estimate V_P for conductance volume measurements in mice (79). This method takes advantage of the relative reactive components of impedance between blood and tissue (80). Despite some theoretical limitations regarding the basic assumptions of field heterogeneity and current leakage, the conductance catheter technique has also been applied to the study of biomechanics in other chambers besides the left ventricle, including the right ventricle (81), right and left atria (82), and aorta (83,84).

Other Pacing Applications. Intracardiac impedance, or transvalvular impedance (TVI), can be used in the assessment of cardiac hemodynamics. This method involves determining the impedance between pacemaker leads in the right atrium and ventricle using a typical dual-chamber pacing configuration. The TVI waveform can be categorized into atrial, valvular, and ventricular components (Fig. 16) (85). Information derived from the atrial component may be useful to identify the loss of electrical capture in the atrium, or the impairment in atrial hemodynamic function associated with supraventricular tachyarrhythmias. The valvular and ventricular components may provide information on the presence, timing, and strength of ventricular mechanical activity (86).

In a study performed by Gasparini et al. (85), the representative TVI tracings (Fig. 16) were recorded from atrial ring to ventricular tip. TVI was measured by application of 64 Hz subthreshold current pulses of 125 μs duration and the amplitude ranging from 15 to 45 μA. The TVI

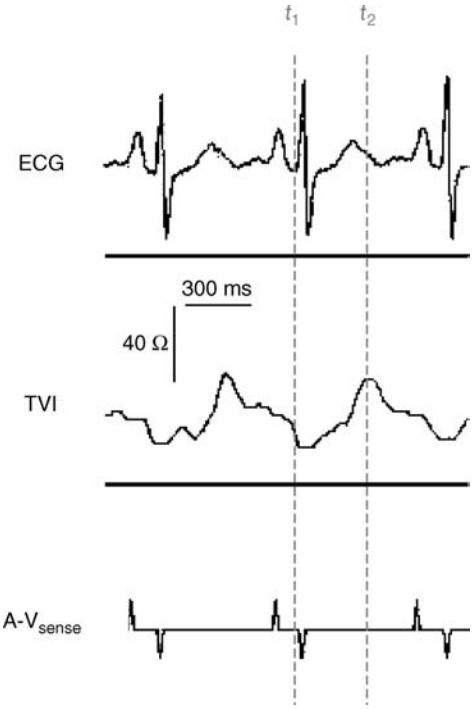


Figure 16. TVI during spontaneous A-V sequential activity. Fiducial points on the TVI waveform used to optimize A-V delay. t_1 corresponds to the end of atrial systole; t_2 corresponds to the end of ventricular ejection (85).

signal was recorded without high-pass filtering to determine the absolute minimum and maximum impedance in each cardiac cycle, which were assumed to reflect the end-diastolic volume and the end-systolic volume, respectively. TVI may represent a useful approach to determine hemodynamic parameters such as stroke volume, ejection fraction, pre-ejection interval, and atrio-ventricular delay. One significant advantage with this technique is that the source and sense leads are of those typically used in pacing systems and offer the advantage of a high signal-to-noise ratio (86). Moreover, the use of this technique has the potential to differentiate atrial from ventricular function that would be paramount if this technique is used for atrioventricular delay optimization (87).

Hematocrit Measurement. Measurement of the resistivity of whole blood has been investigated by a number of researchers, particularly in the area of transthoracic impedance techniques (88–91). A number of investigators have found blood resistivity to be an exponential function of hematocrit (Fig. 17) (15,89–95). These studies have demonstrated a strong correlation between the electrical resistivity of blood at frequencies between 20 to 50 kHz, as the red blood cell is the major resistive component in blood, compared with the relatively conductive plasma. Pop et al. (93) employed a four ring catheter electrode system with narrow electrodes spacing (2 mm center-to-center) to estimate hematocrit in the right atrium of anesthetized pigs. As shown in Fig. 17, good correlation existed between the hematocrit of blood and its electrical resistivity ($r^2 = 0.95$–0.99). Moreover, this study also showed a strong correlation between whole blood viscosity and electrical resistivity.

This interesting observation implies that intracardiac impedance has potential to monitor thrombosis risk in patients with hyperviscosity.

Blood Flow Conductivity Based on Erythrocyte Orientation. The electrical properties of blood are of practical interest in medicine because blood has the highest conductivity of all living tissues (89,96,97). Blood is a heterogeneous suspension of erthrocyctes that have a higher resistivity than the suspending fluid (plasma). The resistivity of blood is a function of the resistivities of plasma, the (fractional) packed-cell volume or hematocrit, and the orientation of the erythrocytes, due to their biconcave shape (98). The orientation of the erythrocytes can be influenced by the viscous forces in flowing blood, resulting in a shear rate-dependent resistivity. In stationary blood, the erythrocytes assume a random distribution while in flowing blood, the plane of the erythrocytes becomes oriented parallel to the axis of flow (99). Thus, minimum resistance occurs when the erythrocytes are oriented in an axial direction, parallel to the stream line. Conversely, maximum resistance occurs when the erythrocytes are oriented in a transverse direction to the stream line (100). The electrical properties of pulsatile blood flow are important when applying transthoracic bioimpedance to estimate cardiac output. In an experiment performed by Katsuyuki et al. (101), erythrocyte orientation, deformation, and axial accumulation caused differences in resistance between flowing and resting blood. Frequency characteristics of blood resistance under pulsatile flow showed that at low pulse rates, the resistance change was minimal, whereas at higher pulse rates, the resistance change increased because the orientation of the erythrocytes cannot follow the rapid changes of pulsatile blood flow. These results suggest that one mechanism of the varying resistance of blood in the aorta during pulsatile blood flow occurs because the orientation of the erythrocyte changes due to shear as a function of heart rate. Therefore, hemodynamic parameters such as cardiac output measured by impedance plethysmography must take into account the anisotropic electrical properties of oriented erythrocytes in blood. Moreover, the resulting resistance of flowing blood depends on the direction of the electrical field applied for impedance measurement and may be affected by the orientation of the erythrocytes during pulsatile flow (101).

REACTIVE APPLICATIONS OF BIOIMPEDANCE

Tissue Impedance

The reactive component of tissue impedance does not contribute significantly to measured impedance when the driving frequency range is less than 1 kHz (8,15). However,

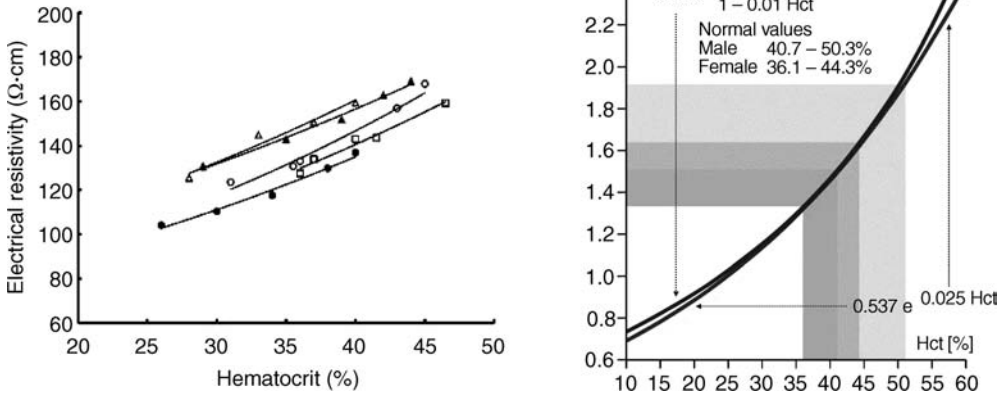


Figure 17. Left figure depicts the correlation between hematocrit of blood and electrical resistivity in five subjects. Right figure depicts a similar correlation between hematocrit of blood and electrical resistivity based on equations by Maxwell–Fricke (upper curve) and Geddes and Sadler (lower curve) (15,93).

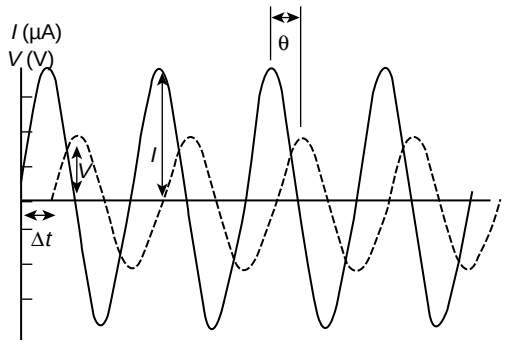


Figure 18. Relationship in phase angle and amplitude for tissue electrical properties. This example shows a capacitive tissue segment since the current waveform (I) leads the voltage waveform (V) by phase angle (θ) (103).

at higher driving current frequencies, the reactive component may contribute more substantially. As different tissues have different reactance, different frequencies may be selected for impedance measurement in order to discriminate various tissues (15,102).

Tissue impedance is characterized by four components: the in-phase component of voltage (V) with respect to the current intensity (I), the tissue resistance (R), and the phase angle (θ). The phase angle represents the time delay between the voltage and current intensity waves due to the capacitance of cell membranes (Fig. 18) (103).

Figure 19 shows cellular tissue structure representing alternating current distribution between a bipolar electrode pair at high and low frequencies. The change in polarity that occurs with AC current causes the cell membrane to charge and discharge at the rate of the applied frequency, and the impedance decreases as a function of increased frequency, because the amount of conducting volume increases through intracellular space. At higher frequencies, the rate of cell membrane charge and discharge becomes such that the effect of the cellular membrane on measured impedance becomes insignificant and the current flows through the intracellular and extracellular space (104).

Capacitance causes the voltage to lag behind the current (Fig. 18), creating a phase shift that is quantified as the angular transformation (θ) of the ratio of reactance to resistance (105). Note that the uniform orientation of cells in a tissue (Fig. 19) can result in anisotropy of electrical properties. That is, impedance will be lower in the longitudinal versus transverse direction of the tissue segment cellular structure (12,18).

The parallel-column model (Fig. 3) must be modified to describe higher frequency applications of bioimpedance in which the capacitive properties of the cell membranes become important. The Cole–Cole plot (Fig. 20) is a useful characterization of the three element RC model that describes the behavior of tissue impedance as a function of frequency (f), impedance (Z), resistance (R), reactance (X_C), and phase angle (ϕ) (103). The real components (R_1 and R_2) can be plotted versus the negative imaginary component of the capacitor (C) with reactance (X_C) in the complex series impedance ($R + jX_C$), with the frequency as a parameter where $j = (\sqrt{-1})$ (15). As the frequency is

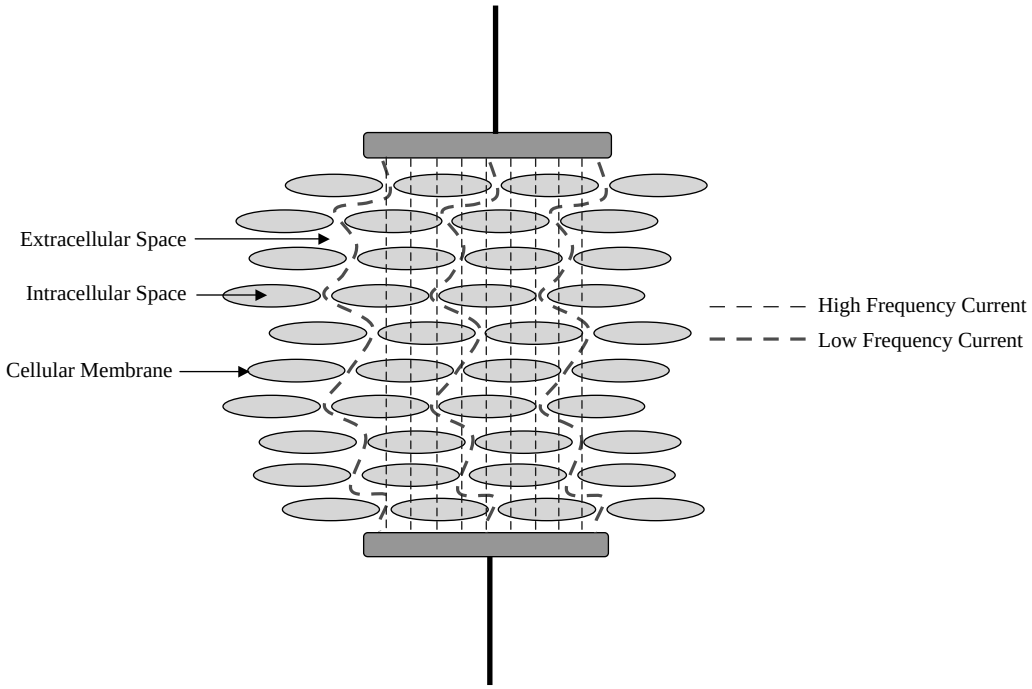


Figure 19. Low and high frequency current distribution is a cellular structure. The low frequency stimulus current flows through the highly conductive extracellular space, whereas the high frequency stimulus current flows through both the extracellular and intracellular space once the reactance of the capacitive cellular membrane is reduced.

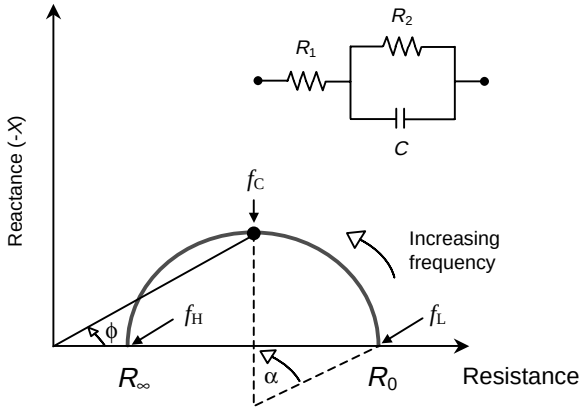


Figure 20. Cole–Cole plot and equivalent tissue impedance circuit. Resistance (abscissa) and reactance (ordinate) plotted as a function of frequency. A three-element electrical equivalent tissue impedance model is shown. At low frequency (f_L), the equivalent circuit is resistive and $R_0 = R_1 + R_2$. As the frequency increases, the phase angle (ϕ) increases until the resistance and reactance are equal at the characteristic frequency of the tissue (f_C). As the frequency increases beyond the characteristic frequency, the reactive element C is reduced to a low impedance and the tissue displays purely resistive properties where $R_\infty = R_1$. The depressed locus at angle α is presumed to represent electrode polarization.

changed between R_0 and R_∞, the impedance will change continually along a curve in the R-X plane. At very low frequencies (f_L), the capacitive component of the system is effectively an open circuit so the reactance is equal to zero and the measured impedance (Z) is purely resistive (R_0). As the frequency increases, reactance (X_C) increases in proportion to resistance causing the phase angle (ϕ) to increase until a maximum angle is reached at the critical (characteristic) frequency (f_C). As shown in Fig. 20, phase angle is positively associated with reactance and negatively associated with resistance (106). Beyond the critical frequency, the reactance begins to decrease in proportion to resistance with increasing frequency and, at very high frequencies (f_H), the capacitive component is essentially short-circuited so the measured impedance is purely resistive at R_∞ (105).

If impedance of a tissue is measure over a broad spectrum, then the resultant impedance Cole–Cole plot can be fit to the three element model or other similar lumped-parameter models. Changes in the model elements can reflect changes in tissue properties due to pathological conditions such as ischemia (see below). In many biologic systems, the center of loci of the plot lies below the real axis and is represented by the angle α, a fixed number between 0 and 1 (2). This behavior can only be modeled by adding an inductive element to the electrical parameter model shown in Fig. 20. However, the physiological interpretation of the inductance is uncertain. Fricke et al. hypothesized that a possible source of this observed inductance might be electrode polarization (107). These investigations demonstrated behavior similar to constant depression angle of electrode polarization. They demonstrated that a frequency-dependent resistance and reactance could mathematically assume a constant depression angle (2). However, the physiologic explanation for $\alpha >$ zero remains controversial. An additional theory related to the origin of the depressed loci is the distribution of time constants in a heterogeneous tissue segment. This distribution could result from variability in cell size or variability in properties of the individual cells (2).

Ischemia Detection. Tissue degradation due to ischemia can alter both the real and reactive components of bioimpedance (40). The dielectric polarization of matter (e.g., myocardial tissue) is given by the dimensionless parameter ε', which is called dielectric permittivity. ε' describes the capacitance increase of a capacitor filled with matter:

$$\varepsilon' = \frac{C}{C_0} \tag{12}$$

where:

C = a capacitor with matter (i.e., cellular structure),
C_0 = vacuum capacitor.

As the dielectric polarization processes are frequency-dependent, they show relaxation phenomena with increasing frequency (108). The relaxation process is defined by the complex dielectric permittivity ε, thus:

$$\varepsilon(\omega) = \varepsilon'(\omega) - i\varepsilon''(\omega) \tag{13}$$

where:

ε' = dielectric permittivity,
ε'' = dielectric loss factor,
$\omega = 2\pi f$,
f = frequency of stimulus current,
i = imaginary unit ($\sqrt{-1}$).

The method of dielectric spectroscopy has been proposed to investigate heart tissue during global ischemia, because the dielectric polarization of matter can be measured by the application of weak electric fields. An electrical circuit model to describe myocardial ischemia, initially developed by Gersing (109) and modified by Schaefer et al. (108), is depicted in Fig. 21. This model can be considered as a variation of the simplified three element physiologic model as shown in Fig. 20. The resistance R_{ext} describes the properties of the extracellular electrolyte, and the resistance R_{int} describes the intracellular cytosol. This model assumes that the transcellular current has to pass the membrane with capacitance C_m and the resistance R_m, through the cytosol, and from cell to cell through the interstitial membranes described by C_{is} or, alternatively, through gap junctions with resistance R_g (108,109). Application of this model enables quantification of the variation of intracellular coupling via gap junctions due to myocardial ischemia (108–111).

The measurement of alterations in impedance spectra with ischemia is often referred to as impedance spectroscopy. Myocardial electrical impedance (MEI), a specific application of impedance spectroscopy, has been shown to

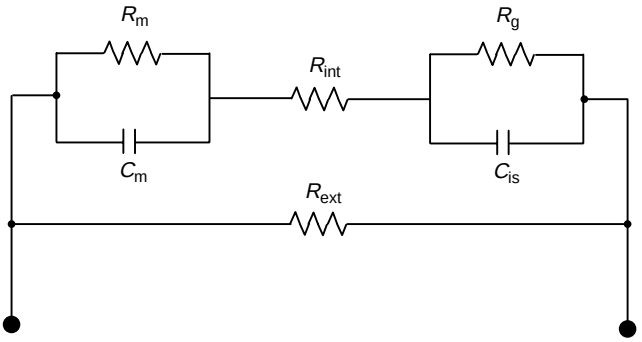


Figure 21. Electronic equivalent model of heart tissue. Electrical elements consist of the intracellular (R_{int}) and extracellular resistance (R_{ext}), cellular membrance capacitance (C_m) and resistance (R_m), cell-to-cell interstitial membrane capacitance (C_{is}), and gap junction resistance (R_g).

identify localized and global myocardial tissue in various disease states in *in vitro* and *in vivo* experimental models (112).

Recently, MEI has been used in conjunction with electrocardiogram (ECG) ST-segment deviations to assess the magnitude of the ischemic region of the myocardium (103,112–116).

Injury currents, secondary to myocardial ischemia result in ST-segment displacements in the ECG of patients with myocardial ischemia (117). Injury currents deriving from resting depolarization in ischemic myocardial cells are associated with slow conduction through the myocardium. The mechanisms by which these injury currents correlate with the impedance spectroscopy alterations in the ischemic myocardial tissue are well described (103,108,109,112–117). Figure 22 depicts a segment of the myocardium with subendocardial and transmural ischemic tissue. Blood flow through the heart is interrupted during ischemia, and the tissue undergoes progressive changes leading to irreversible loss of its viability (108). Transmural ischemia causes ST-segment elevation and subendocardial ischemia causes ST-segment depression (117). The electrocardiographic differences between transmural and subendocardial ischemia are clinically important. Supply ischemia, as occurs following total interruption of flow through a coronary artery supplying a large area of the left ventricle, typically causes ST-segment elevation (63). In contrast, demand ischemia, as occurs during a stress test, begins in the subendocardial regions of the left ventricle and causes ST-segment depression (117).

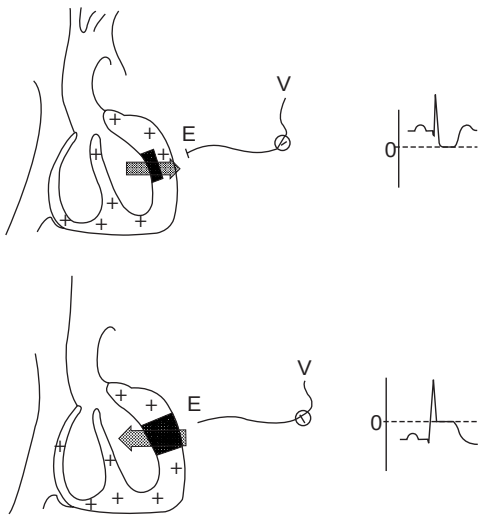


Figure 22. Ischemic regions of myocardial tissue and corresponding ST-segment. Subendocardial ischemia (top) with depressed ST-segment. Transmural ischemia (bottom) with elevated ST-segment (117).

The mechanism by which MEI changes with ischemia is not certain, but may well be associated with ultrastructural changes or cellular biochemical changes that occur in the myocardial tissue similar to those viewed by ST-segment deviations (113). The increase in MEI may result from reductions in the conductive fluid volume in the affected region of the myocardium (113). Gap junctions play a critical role in the propagation of electrical impulse in the heart, and its conductivity has been shown to be reduced and eventually abolished during ischemia and rapidly restored during reperfusion (103). Thus, gap junction closure is a reasonable hypothesis to explain observed impedance changes with ischemia. The intraischemic variation of intracellular and extracellular coupling is one possible explanation for the observed impedance changes of the dielectric frequency spectrum (108).

As MEI correlates with myocardial tissue viability (118,119), the measure has several important potential monitoring applications. Intraoperatively, MEI could be used to detect ischemia in aortic or myocardial tissue during cardiopulmonary bypass surgery as an early indication of damage. Following cardiopulmonary bypass, MEI could be used to assess reperfusion afforded by the new grafts. MEI could also aid in drug titration after cardiac surgery as well as to chronically monitor tissue perfusion with implantable devices such as pacemakers or cardioverter-defibrillators, or with patients whom have received a heart transplant (12,113,120).

In a study performed by Howie et al. (113), acute ischemia was induced in anesthetized dogs via left anterior descending (LAD) coronary artery occlusion for randomly assigned periods of 15, 30, 45, 60, or 120 min. MEI was simultaneously recorded using ventricular pacing leads sutured into the exposed heart tissue. As shown in Fig. 23, MEI increased immediately after LAD coronary artery occlusion and returned to baseline following reperfusion. A statistically significant increase occurred from baseline impedance when compared at 64, 68, 72, 76, and 80 min (113). This intracardiac technique used by Howie et al. suggested other possible applications for MEI with implantable devices and intracardiac pacing/monitoring leads. However, further development in the direction of optimal electrode placement to isolate the targeted tissue region and obtain the highest quality data for diagnosis of tissue alteration is warranted (12).

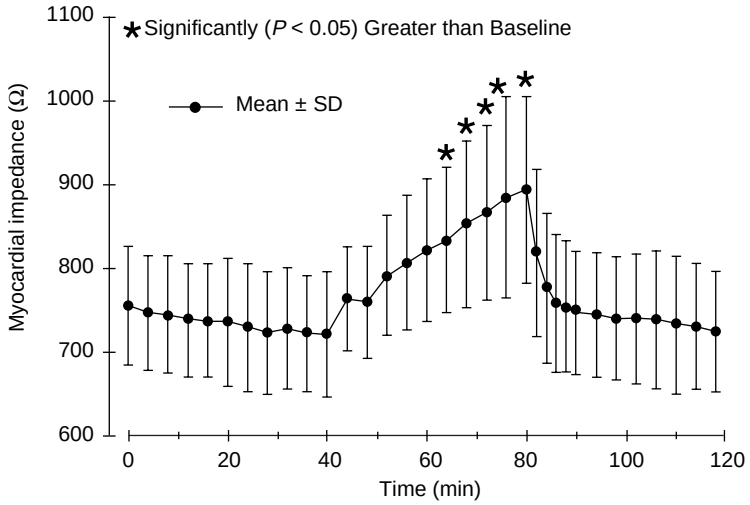


Figure 23. Change in myocardial impedance during LAD coronary artery occlusion (113).

Dialysis. Whole-body bioimpedance spectroscopy has been proposed by several investigators for measuring extracellular (ECW) and intracellular (ICW) water volumes in dialysis patients in order to assess nutritional status and to monitor hydration (121–125). An adequate assessment of body water compartments is crucial in dialysis patients because overhydration and underhydration are often difficult to detect and may result in severe morbidity in this population (126). Despite the continuous progress in the delivery of renal replacement therapy, mortality in patients on maintenance dialysis remains higher than in the general population (127).

During acute volume overload, most of the extra fluid collects in the ECW not the ICW. At very low frequencies, current only penetrates the ECW because the cell membrane acts as a capacitor and the impedance becomes equal to the ECW resistance (see Fig. 19). At very high frequencies, the injected current penetrates both the ECW and the ICW, and the impedance represents the total body water (TBW) resistance (125). Several investigators (128–131) have used single- and multiple-frequency impedance to monitor fluid shifts during hemodialysis. However, when attempting to determine precise fluid volumes from the measured impedance, difficulties occur due to the complex geometry of the human body and electrical inhomogeneity of nonconducting elements such as bone and fat (125). Signal processing methods to account for these aforementioned difficulties are described in the literature (104,124,125,132).

Whole-body bioelectrical impedance measurements typically apply single (e.g., 50 kHz) (133) or multifrequency (e.g., 5 to 1000 kHz) alternating currents applied via cutaneous electrodes placed on the hands and feet with more proximal electrodes uses for voltage measurements (126). The precise method for calculation of body fluid volumes depends on whether the single-frequency or multiple-frequency method is applied. The single-frequency method often uses an empirically derived regression formula to assess TBW, whereas the multiple-frequency method predicts the volume of TBW and ECW from a general mixture theory, assuming specific resistance values for ECW and ICW (104,126,134). Moreover, the contribution of body weight, which is strongly related to ECW and TBW, is greater in the regression approach compared with the mixture approach (126).

Although reliable measurements of fluid content in dialysis patients have been reported (121–131), uncertainty remains regarding the agreement of whole-body bioimpedance in dialysis patients with tracer dilution techniques, which are considered the gold standard methods (126). One explanation for the lack of satisfactory agreement between techniques is that whole-body bioimpedance techniques consider the body as a multiple conductive cylinder model (e.g., arms, legs, trunk) connected in series (Fig. 24). With conductors connected in series, conductors with the smallest cross-sectional area (e.g., extremities) will determine most of the resistance, whereas the component with the largest cross-sectional area (e.g., trunk) will have minimal contribution to the resistance

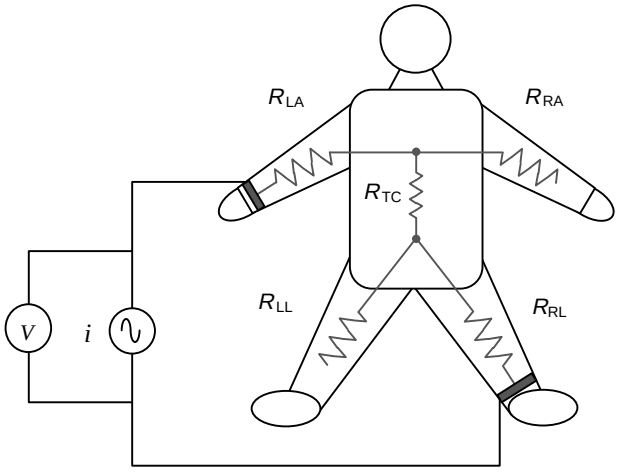


Figure 24. Whole-body impedance measurement technique. Total Conductance (C_T) = Left Arm Conductance ($1/R_{LA}$) + Thor-Thoracic Cavity Conductance ($1/R_{TC}$) + Right Leg Conductance ($1/R_{RL}$).

although it contains a significant amount of body water (126). However, assessment of the sum of segmental bioelectrical impedance analysis measurements, which take into account resistance of the extremities and the trunk independently, have been shown to detect changes in trunk water more accurately (135). A seminal study performed by Patterson et al. (136) used multiple linear regression analysis combining data measured independently from the arms, legs, and trunk correlated with weight change on patients undergoing hemodialysis, gave a correlation coefficient of 0.87, whereas the correlation coefficient from measurements between just the wrist and ankle was 0.64.

Pulmonary Edema Detection. Patients developing pulmonary edema initially accumulate fluid in the interstitial spaces of the lung. As the condition progresses, fluid ultimately accumulates in the alveoli. To accurately measure pulmonary fluid status, the different bioelectric properties of blood, lung tissue, and extravascular fluid must be considered, and an impedance parameter not influenced by the patient's geometry should be used (137). Thus, using a dual-frequency measurement of thoracic impedance, an impedance ratio can be calculated that represents the ratio between intracellular and extracellular water. This ratio, therefore, changes as a result of the fluid shift caused by edema formation. As the low frequency current only passes through the extracellular resistance, the measured low frequency impedance (Z_{LF}) over a specified thoracic length equals the total extracellular resistance. As the frequency is increased, current is divided over the intracellular and extracellular compartments. Therefore, the measured high frequency impedance (Z_{HF}) over a specified thoracic length equals the parallel equivalent of intracellular and extracelluar impedance. Thus, a dual-frequency impedance ratio that represents the intracellular/extracellular impedance fraction can be defined by Z_{HF}/Z_{LF}. As pulmonary fluid accumulates in the extracellular space, the impedance ratio increases (137).

BIBLIOGRAPHY

Cited References

1. Grimnes S, Martinsen O. Bioimpedance and Bioelectricity Basics. London: Academic Press; 2000. p 313–320.
2. Ackmann J, Seitz M. Methods of complex impedance measurements in biologic tissue. CRC Crit Rev Biomed Eng 11(4):281–311.
3. Schwan HP. The bioimpedance field: Some historical observations. Proceedings of the 4th International Conference on Electrical Bioimpedance. 1995; 1–4.
4. Nyboer J, et al. Radiocardiograms: Electrical impedance changes of the heart in relation to electrocardiograms and heart sounds. J Clin Invest 1940; 19:963.
5. Nyboer J. Electrical impedance plethysmography: A physical and physiologic approach to peripheral vascular study. Circulation 1950;2:811–821.
6. Kubicek W, et al. Development and evaluation of an impedance cardiac output system. Aerospace Med 1966;37:1208–1212.
7. Djordjevich L, Sadove M. Basic principles of electrohemodynamics. J Biomed Eng 1981; 3:25–33.
8. Baan J, et al. Continuous measurement of left ventricular volume in animals and humans by conductance catheter. Circulation 1984;70(5):812–823.
9. Min M, Parve T, Kink A. Thoracic impedance as a basis for pacing control. Ann NY Acad Sci 1999;873: 155–166.
10. Ellenbogen K, Wood M. Cardiac Pacing and ICD's. 3rd ed. Malden (MA): Blackwell Science; 2002. p 106–109.
11. Kindermann M, et al. Optimizing the AV delay in DDD pacemaker patients with high degree AV block: Mitral valve doppler versus impedance cardiography. PACE 1997;20(10 pt 1):2453–2462.
12. Steendijk P, et al. The four-electrode resistivity technique in anisotropic media: Theoretical analysis and application on myocardial tissue in vivo. IEEE Trans Biomed Eng 1993;40(11):1138–1147.
13. Nyboer J. Electrorheometric properties of tissues and fluids. Ann NY Acad Sci 1970;170(2):410–420.
14. Geddes LA, Hoff HE. The measurement of physiological events by electrical impedance, a review. Am J Med Electron 1964; 14:16–27.
15. Malmivuo J, Plonsey R. Bioelectromagnetism. In: Principles and Application of Bioelectric and Biomagnetic Fields. United Kingdom: Oxford Univeristy Press; 1995. p 141, 405–419.
16. Hettrick DA, Battocletti JH, Ackmann JJ, Linehan JH, Warltier DC. Effects of physical parameters on the cylindrical model for conductance volume measurement. Ann Biomed Eng 1997;24:126–134.
17. Hettrick D, et al. Finite element model determination of correction factors used for measurement of aortic diameter via conductance. Ann Biomed Eng 1999;27(2): 151–159.
18. Roberts DE, Hersh LT, Scher AM. Influence of cardiac fiber orientation on wavefront voltage, conduction velocity, and tissue resistivity in the dog. Circ Res 1979; 44:701–712.
19. Patterson R, et al. Development of an electrical impedance plethysmography system to monitor cardiac output. Proceedings of the First Annual Rocky Mountain Bioengineering Symposium, 1964: 56–71.
20. Sramek B, Rose D, Miyamoto A. A stroke volume equation with a linear base impedance model and its accuracy, as compared to thermodilution and magnetic flowmeter techniques in humans and animals. Proceedings of the Sixth International Conference on Electrical Bioimpedance, Zadar, Yugoslavia, 1983: 38.
21. Milnor WR. Hemodyanmics. Baltimore: Williams & Wilkins Co.; 1982. p 155.
22. Statistical Bulletin. Metropolitan Life Foundation, January-June 1983. p 64.
23. Quail A, et al. Thoracic resistivity for stroke volume calculation by impedance cardiography. J Appl Physiol 1981;50:191.
24. Albert N, et al. Equivalence of bioimpedance and thermodilution in measuring cardiac output in hospitalized patients with advanced, decompensated chronic heart failure. Am J Crit Care 2004;13:469–479.
25. Van DeWater J, et al. Impedance cardiography: The next vital sign technology? Chest 2003;123: 2028–2033.
26. Drazner M, et al. Comparison of impedance cardiography with invasive hemodynamic measurments in patients with heart failure secondary to ischemic or nonischemic cardiomyopathy. Am J Cardiol 2002;89:993–995.
27. Sageman W, Riffenburgh R, Spiess B. Equivalence of bioimpedance and thermodilution in measuring cardiac index

after cardiac surgery. J Cardiothroracic Vasc Anesth 2002; 16:8–14.
28. Yung G, et al. Comparison of impedance cardiography to direct Fick and thermodilution cardiac output determination in pulmonary arterial hypertension. Congest Heart Fail 2004;10(Suppl 2):7–10.
29. Patterson RP. Possible technique to measure ventricular volume using electrical impedance measurements with an oesophageal electrode. Med Biol Eng Comput 1987;25:677–679.
30. Hettrick DA, et al. Correlation of esophageal conductance measurements with aortic and left ventricular diameters and stroke volume. IEEE Trans Biomed Eng 2000;47:559–564.
31. Patterson R. Handbook of Biomedical Engineering. Bioelectric Impedance Measurements. Boca Raton, FL: CRC Press; 1995. p 1223–1230.
32. Ebert T, et al. The use of thoracic impedance for determining thoracic blood volume changes in man. Aviation Space Environ Med 1986 57:49–53.
33. Mancini R, et al. Cardiac output and contractility indices: Establishing a standard in response to low-to-moderate level exercise in healthy men. Arch Phys Med Rehabil 1979;60:567–573.
34. Levett J, Replogle R. Thermodilution cardiac output: A critical analysis and review of the literature. J Surg Res 1979;27:392–404.
35. Engoren M, Barbee D. Comparison of cardiac output determined by bioimpedance, thermodilution, and the Fick method. Am J Crit Care 2005;14(1):40–45.
36. Imhoff M, Lehner J, Lohlein D. Noninvasive whole-body electrical bioimpedance cardiac output and invasive thermodilution cardiac output in high-risk surgical patients. Crit Care Med 2000;28:2812–2818.
37. Cotter G, et al. Accurate, Noninvasive continuous monitoring of cardiac output by whole body electrical bioimpedance. Chest 2004;125:1431–1440.
38. Raaijmakers E, et al. A meta-analysis of published studies concerning the validity of thoracic impedance cardiography. Ann NY Acad Sci 1999;873: 121–134.
39. Patterson R, Witsoe D. Impedance stroke volume compared with dye and electromagnetic flowmeter values during drug induced inotropic and vascular changes in dogs. Ann NY Acad Sci 1999;873:143–148.
40. Min M, Ollmar S, Gersing E. Electrical impedance and cardiac monitoring: Technology, potential and applications. Int J Bioelectromag 2003;5(1):53–56.
41. Kim D, et al. Origins of the impedance change in impedance cardiography by a three-dimensional finite element model. IEEE Trans Biomed Eng 1988 ;35(12): 993–1000.
42. Wang L, Patterson R. Multiple sources of the impedance cardiogram based on 3-D finite difference human thorax models. IEEE Trans Biomed Eng 1995;42(4): 141–148.
43. Wang X, et al. An impedance cardiography system: A new design. Ann Biomed Eng 1989;17: 535–556.
44. Summers R, Kolb J, Woodward L. Differentiating systolic from diastolic heart failure using impedance cardiography. Academ Emerg Med 1999;6(7):693–699.
45. Lababidi Z, et al. The first derivative thoracic impedance cardiogram. Circulation 1970;41(4): 651–658.
46. Osypka M, Berstein D. Electrophysiologic principles and theory of stroke volume determination by thoracic electrical bioimpedance. Non-invasive monitoring using thoracic bioimpedance. AACN Clin Issues 1999;10(3): 385–399.
47. Christov I. Dynamic powerline interference subtraction from biosignals. J Med Eng Technol 2000;24(4):169–172.
48. Wang X, Sun H, Van DeWater J. An advanced signal processing technique for impedance cardiography. IEEE Trans Biomed Eng 1995;42(2):224–230.
49. Barros A, Yoshizawa M, Yasuda Y. Filtering noncorrelated noise in impedance cardiography. IEEE Trans Biomed Eng 1995;42(3):324–327.
50. Eiken O, Segerhammer P. Elimination of breathing artifacts from impedance cardiograms at rest and during exercise. Med Biol Eng Comput 1988;13–16.
51. Raza S, Patterson R, Wang L. Filtering respiration and low-frequency movement artifacts from the cardiogenic electrical impedance signal. Med Biol Eng Comput 1992; 556–561.
52. Abdelhammed A, et al. Noninvasive hemodynamic profiles in hypertensive subjects. Am J Hypertens 2005;18:51S–59S.
53. Ventura H, Taler S, Strobeck J. Hypertension as a hemodynamic disease: The role of impedance cardiography in diagnostic, prognostic, and therapeutic decision making. Am J Hypertens 2005;18:26S–43S.
54. Alfie J, Galarza C, Waisman G. Noninvasive hemodynamic assessment of the effect of mean arterial pressure on the amplitude of pulse pressure. Am J Hypertens 2005;18:60S–64S.
55. Taler S, Textor S, Augustine J. Resistant hypertension: Comparing hemodynamic management to specialist care. Hypertension 2002;39:982–988.
56. Sharman D, Gomes C, Rutheford J. Improvement in blood pressure control with impedance cardiograph-guided pharmacologic decisions making. Congest Heart Fail 2004;10: 54–58.
57. Eugene M, et al. Assessment of the optimal atrio-ventricular delay in DDD paced patients by impedance plethysmography. Eur Heart J 1989;10:250–255.
58. Ovsyshcher I, et al. Measurements of cardiac output by impedance cardiography in pacemaker patients at rest: Effects of various atrioventricular delays. JACC 1993; 21(3):761–767.
59. Tse H, et al. Impedance cardiography for atrioventricular interval optimization during permanent left ventricular pacing. PACE 2003;26(Pt II):189–191.
60. Wolf GK, Arnold JH. Noninvasive assessment of lung volume: Respiratory inductance plethysmography and electrical impedance tomography. Crit Care Med 2005;33(3): S163–S169.
61. Coulombe N, et al. A parametric model of the relationship between EIT and total lung volume. Physiol Meas 2005;26(4):401–411.
62. Zhang J, Patterson RP. EIT images of ventilation: What contributes to the resistivity changes? Physiol Meas 2005; 26(2):S81–S92.
63. Edd JF, Horowitz L, Rubinsky B. Temperature dependence of tissue impedivity in electrical impedance tomography of cryosurgery. IEEE Trans Biomed Eng 2005;52(4): 695–701.
64. Xiao C, Lei Y. Analytical solutions of electric potential and impedance for a multilayered spherical volume conductor excited by time-harmonic electric current source: Application in brain EIT. Phys Med Biol 2005;750(11): 2663–2674.
65. Clay MT, Ferree TC. Weighted regularization in electrical impedance tomography with applications to acute cerebral stroke. IEEE Trans Med Imag 2002; 21(6):629–637.

66. Zlochiver S, Rosenfeld M, Shimon A. Contactless bio-impedance monitoring technique for brain cryosurgery in a 3D head model. Ann Biomed Eng 2005;33(5):616–625.
67. Geselowitz D. An application of electrocardiographic lead theory to impedance plethysmography. IEEE Trans Biomed Eng 1971;18(1):38–41.
68. Kauppinen P, et al. Application of computer modelling and lead field theory in developing multiple aimed impedance cardiography measurments. J Med Eng Technol 1999;23(5):169–177.
69. Kauppinen P, et al. Lead field theoretical approach in bioimpedance measurements: Towards more controlled measurment sensitivity. Ann NY Acad Sci 1999; 135–142.
70. Wang L, Lahtinen S, Lentz L, et al. Feasibility of using and implantable system to measure thoracic congestion in an ambulatory chronic heart failure canine model. PACE 2005; 28:404–411.
71. Pomerantz M, et al. Transthoracic electrical impedance for the early detection of pulmonary edema. 1969;66:260–268.
72. Fein A, et al. Evaluation of transthoracic electrical impedance in the diagnosis of pulmonary edema. Circulation 1979;60:1156–1160.
73. Gotshall R, Davrath L. Bioelectric impedance as an index of thoracic fluid. Aviation Space Environ Med 1999;70(1):58–61.
74. Yu C, et al. Intrathoracic impedance monitoring in patients with heart failure. Correlation with fluid status and feasibility of early warning preceding hospitalization. Circulation 2005;112:841–848.
75. Steendijk P, et al. Pressure-volume measurements by conductance catheter during cardiac resynchronization therapy. Eur Heart J Suppl 2004;6(Suppl D): D35–D42.
76. Baan J, Van der Velde E, Steendijk P. Ventricular pressure-volume relations in vivo. Eur Heart J 1992;13(Suppl E):2–6.
77. Segers P, et al. Conductance catheter based assessment of arterial input impedance, arterial function, and ventricular-vascular interaction in mice. Am J Physiol Heart Circu Physiol 2005;288:H1157–H1164.
78. Szwarc RS, Laurent D, Allegrini PR, Ball HA. Conductance catheter measurement of left ventricular volume: evidence for nonlinearity within cardiac cycle. Am J Physiol 1995;268: H1490–H1498.
79. Georgakopoulus D, Kass DA. Estimation of parallel conductance by dual-frequency conductance catheter in mice. Am J Physiol Heart Circu Physiol 2000;279:H443–H450.
80. Gawne TJ, Gray KS, Goldstein RE. Estimating left ventricular offset volume using dual frequency conductance catheters. J Appl Physiol 1987;63:872–876.
81. Nicolosi AC, Hettrick DA, Warltier DC. Assessment of right ventricular function in swine using sonomicrometry and conductance. Ann Thorac Surg 1996;61:1281–1387.
82. Schwartzman D, et al. Atrial pacing lead location alters left atrial-ventricular mechanical coupling relationships independent of AV delay in humans: A dual-chamber pressure-volume analysis. Heart Rhythm 2005;2:S85.
83. Hettrick DA, et al. In vivo measurement of real time aortic segmental volume using the conductance catheter. Ann Biomed Eng 1998;26:431–440.
84. Kornet L, et al. Conductance method for the measurement of cross-sectional areas of the aorta. Ann Biomed Eng 1999;27:141–150.
85. Gasparini G, et al. Rate-responsive pacing regulated by cardiac hemodynamics. Europace 2005; 7:234–241.
86. Di Gregorio F, et al. Transvalvular impedance (TVI) recording under electrical and pharmacological cardiac stimulation. PACE 1996;19(Pt.II):1689–1693.
87. Salo R. Application of impedance volume measurement to implantable devices. Int J Bioelectromagn 2003;5(1): 57–60.
88. Fricke H, Morse S. The electrical resistance and capacity of blood for frequencies between 800 and 4.5 MHz. J Gen Physiol 1926;9:153–167.
89. Geddes L, Sadler C. The specific resistance of blood at body temperature. IEEE Trans Biomed Eng 1973;20: 336–339.
90. Hill D, Thompson F. The effect of hematocrit on the resistivity of human blood at 37 degrees celsius and 100 kHz. Med Biol Eng 1975;March:182–186.
91. Mohapatra S, Costeloe K, Hill D. Blood resistivity and its implications for the calculation of cardiac output by the thoracic electrical impedance technique. Intens Care Med 1977;3:63–67.
92. Fuller H. The electrical impedance of plasma: A laboratory simulation of the effect of changes in chemistry. Ann Biomed Eng 1991;19:123–129.
93. Pop G, et al. Catheter based impedance measurements in the right atrium for continuously monitoring hematocrit and estimating blood viscosity changes. An in vivo feasibility study in swine. Biosens Bioelectron 2004;19:1685–1693.
94. Geddes L, Sadler C. The specific resistance of blood at body temperature. Med Biol Eng 1973;11(5):336–339.
95. Fricke H. A mathematical treatment of the electric conductivity and capacity of disperse systems. Physiol Rev 1924;4:575–587.
96. Sigman E, Kolin A, Katz L. Effect of motion on the electrical conductivity of blood. Am J Physiol 1937;118:708.
97. Gollan F, Namon R. Electrical impedance of pulsatile blood flow in rigid tubes and in isolated organs. Ann NY Acad Sci 1970;170(2):568–576.
98. Visser K. Electric properties of flowing blood and impedance cardiography. Ann Biomed Eng 1989;17: 463–473.
99. Peura R, et al. Influence of erythrocyte velocity on impedance plethysmographic measurements. Med Biol Eng Comput 1978;16:147–154.
100. Tanaka K, et al. The impedance of blood: The effects of red cell orientation and its application. Japan J Med Electron Biol Eng 1970;8:14–21.
101. Katsuyuki S, Hiroshi K. Electrical characteristics of flowing blood. IEEE Trans Biomed Eng 1979;26(12):686–695.
102. Lozano A, Rossell J, Pallas-Areny R. Two frequency impedance plethysmograph: Real and imaginary parts. Med Biol Eng Comput 1990;28(1):38–42.
103. Padilla F, et al. Protection afforded by ischemic preconditioning is not mediated by effects on cell-to-cell electrical coupling during myocardial ischemia reperfusion. Am J Heart Circ Physiol 2003;285:H1909–H1916.
104. De Lorenzo A, et al. Predicting body cell mass with bioimpedance by using theoretical methods: A technological review. J Appl Physiol 1997;82(5):1542–1558.
105. Baumgartner R, Chumlea W, Roche A. Bioelectric impedance phase angle and body composition. Am J Clin Nutr 1988; 48:16–23.
106. Barnett A, Bango S. The physiological mechanisms involved in the clinical measure of phase angle. Am J Physiol 1936; 114:366–382.
107. Fricke H. The theory of electrolyte polarization. Phil Mag 1932;14:310.
108. Schaefer M, et al. The complex dielectric spectrum of heart tissue during ischemia. Bioelectrochemistry 2002;58:171–180.

109. Gersing E. Impedance spectroscopy on living tissue for determination of the state of organs. Bioelectrochem Bioenerget 1998;45:145–149.
110. Owens L, et al. Correlation of ischemia-induced extracellular and intracellular ion changes to cell-to-cell electrical uncoupling in isolated blood-perfused rabbit hearts. Circulation 1996;94:10–13.
111. Schafer M, Gebhard Gersing E. Characterization of organ tissue during the transition between life and death: Cardiac and skeletal muscle. Med Biol Eng Comput 1999;37:100–101.
112. Dzwonczyk R, et al. Myocardial electrical impedance responds to ischemia and reperfusion in humans. IEEE Trans Biomed Eng 2004;51(12):2206–2209.
113. Howie M, Dzwonczyk R, McSweeney T. An evaluation of a new two-electrode myocardial electrical impedance monitor for detecting myocardial ischemia. Anesthesia Analgesia 2001;92:12–18.
114. Sezer M, et al. New support for clarifying the relation between ST segment resolution and microvascular function: Degree of ST segment resolution correlates with the pressure derived collateral flow index. Heart 2004;90: 146–150.
115. Leung J, et al. Automated electrocardiograph ST segment trending monitors: Accuracy in detecting myocardial ischemia. Anesthesia Analgesia 1998;87:4–10.
116. Leung J, et al. Electrocardiographic ST-segment changes during acute, severe isovolemic hemodilution in humans. Anesthesiology 2000;93:1004–1010.
117. Katz A. Physiology of the Heart. 2nd ed. New York: Raven Press; 1992. p 609–637.
118. Garrido H, et al. Bioelectrical tissue resistance during various methods of myocardial preservation. Ann Thorac Surg 1983;36:143–151.
119. Gebhard M, et al. Impedance spectroscopy: A method for surveillance of ischemia tolerance of the heart. Thorac Cardiovasc Surg 1987;35:26–32.
120. Mueller J, et al. Electric impedance recording: A noninvasive method of rejection diagnosis. J Extra Corpor Technol 1992;23:49–55.
121. Matthie J, et al. Development of commercial complex bioimpedance spectroscopic system for determining intracellular and extracellular water volumes. Proceedings of the 8th International Conference on Electrical Bioimpedance, Kupio, Finland, 1992. p 203–205.
122. Van Loan M, et al. Use of bioimpednace spectroscopy to determine extracellular fluid, intracellular fluid, total body water, and fat-free mass. In: Human body composition: In vivo methods, models and assessment. New York: Plenum; 1993. p 67–70.
123. Van Marken Lichtenbelt W, et al. Validation of bioelectric impedance measurements as a method to estimate body water compartments. Am J Clin Nutr 1994;60: 159–166.
124. Van Loan M, et al. Fluid changes during pregnancy: Use of bioimpedance spectroscopy. J Appl Physiol 1995;27:152–158.
125. Jaffrin M, et al. Continuous monitoring of plasma, interstitial, and intracellular fluid volumes in dialyzed patients by bioimpedance and hematocrit measurements. ASAIO J 2002;48:326–333.
126. Cox-Reijven P, et al. Role of bioimpedance spectroscopy in assessment of body water compartments in hemodialysis patients. Am J Kidney Dis 2001; 38(4):832–838.
127. Mancini A, et al. Nutritional status in hemodialysis patients and bioimpedance vector analysis. J Renal Nutrition 2003;13(3):199–204.
128. DeVries P, et al. Measurement of transcellular fluid shifts during hemodialysis. Med Biol Eng Comput 1989;27:152–158.
129. Sinning W, et al. Monitoring hemodialysis with bioimpedance: What do we really measure? ASAIO J 1993;39:M584–M589.
130. Scanferla F, et al. On-line bioelectric impedance during hemodialysis: Monitoring of body fluids and cell membrane status. Nephrol Dial Transplant 1990; 5(Suppl 1):167–170.
131. Jaffrin M, et al. Extracellular and intracellular fluid volume during dialysis by multifrequency impedancemetry. ASAIO J 1996;42:M533–M537.
132. Hanai T. Electrical properties of emulsions. In: Emulsions Science. London: Academic Press; 1968. p 354–477.
133. Foley K, et al. Use of single-frequency bioimpedance at 50 kHz to estimate total body water in patients with multiple organ failure and fluid overload. Crit Care Med 1999;27(8):1472–1477.
134. Ward L, Elia M, Cornish B. Potential errors in the application of mixture theory to multifrequency bioelectrical impedance analysis. Physiol Meas 1998;19:53–60.
135. Zhu F, et al. Estimation of body fluid changes during peritoneal dialysis by segmental bioimpedance analysis. Kidney Int 2000;57:299–306.
136. Patterson R, et al. Measurement of body fluid volume change using multisite impedance measurements. Med Biol Eng Comput 1988;26:33–37.
137. Raaijmakers E, et al. Estimation of non-cardiogenic pulmonary edema using dual-frequency electrical impedance. Med Biol Eng Comput 1998;36:461–466.

Further Reading

Cole KS, Cole RH. Dispersion and absorption in dielectrics. J Chem Phys 1941;9:341–351.

See also ELECTROCARDIOGRAPHY, COMPUTERS IN; EXERCISE STRESS TESTING; FLOWMETERS, ELECTROMAGNETIC; IMPEDANCE PLETHYSMOGRAPHY; NEONATAL MONITORING; PHONOCARDIOGRAPHY.

BIOINFORMATICS

ALI ABBAS
LEI LIU
University of Illinois
Urbana, Illinois

INTRODUCTION

The past two decades have witnessed revolutionary changes in biomedical research and biotechnology and an explosive growth of biomedical data. High throughput technologies developed in automated DNA sequencing, functional genomics, proteomics, and metabolomics enable production of such high volume and complex data that the data analysis becomes a big challenge. Consequently, a promising new field, bioinformatics has emerged and is growing rapidly. Combining biological studies with computer science, mathematics, and statistics, bioinformatics develops methods, solutions, and software to discover patterns, generate models, and gain insight knowledge of complex biological systems.

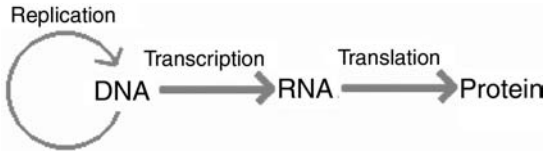


Figure 1. Central dogma of molecular biology.

Before bioinformatics is discussed further, a brief review of the basic concepts in molecular biology, which are the foundations for bioinformatics studies, is provided. The genetic information is coded in DNA sequences. The physical form of a gene is a fragment of DNA. A genome is the complete set of DNA sequences that encode all the genetic information for an organism, which is often organized into one or more chromosomes. The genetic information is decoded through complex molecular machinery inside a cell composed of two major parts, transcription and translation, to produce functional protein and RNA products. These molecular genetic processes can be summarized precisely by the central dogma shown in Fig. 1. The proteins and active RNA molecules combined with other large and small biochemical molecules, organic compounds, and inorganic compounds form the complex dynamic network systems that maintain the living status of a cell. Proteins form complex 3D structures that carry out functions. The 3D structure of a protein is determined by the primary protein sequence and the local environment. The protein sequence is decoded from the DNA sequence of a gene through the genetic codes as shown in Table 1. These codes have been shown to be universal among all living forms on earth.

The high throughput data can be generated at many different levels in the biological system. The genomics data are generated from the genome sequencing that deciphers the complete DNA sequences of all the genetic information in an organism. We can measure the mRNA levels using microarray technology to monitor the gene expression of all the genes in a genome known as transcriptome. Proteome is the complete set of proteins in a cell at a certain stage, which can be measured by high throughput 2D gel electrophoresis and mass spectrometry. We also can monitor all the metabolic compounds in a cell known as metabolome in a high throughput fashion. Many new terms ending with "ome" can be viewed as the complete set of entities in a cell. For example, the "interactome" refers to the complete set of protein-protein interactions in a cell.

Bioinformatics is needed at all levels of high throughput systematic studies to facilitate the data analysis, mining, management, and visualization. But more importantly, the major task is to integrate data from different levels and prior biological knowledge to achieve system-level understanding of biological phenomena. As bioinformatics touches on many areas of biological studies, it is impossible to cover every aspect in a short chapter. In this chapter, the authors will provide a general overview of the field and focus on several key areas, including sequence analysis, phylogenetic analysis, protein structure, genome analysis, microarray analysis, and network analysis.

Sequence analysis often refers to sequence alignment and pattern searching in DNA and protein sequences. This area can be considered classic bioinformatics, which can be dated back to 1960s, long before the word bioinformatics appeared. It deals with the problems such as how to make an optimal alignment between two sequences and how to

Table 1. The Genetic Code

	Second Position				
First Position	T	C	A	G	Third Position
T	TTT Phe [F]	TCT Ser [S]	TAT Tyr [Y]	TGT Cys [C]	T
	TTC Phe [F]	TCC Ser [S]	TAC Tyr [Y]	TGC Cys [C]	C
	TTA Leu [L]	TCA Ser [S]	TAA *Stop*[end]	TGA *Stop*[end]	A
	TTG Leu [L]	TCG Ser [S]	TAG *Stop*[end]	TGG Trp [W]	G
C	CTT Leu [L]	CCT Pro [P]	CAT His [H]	CGT Arg [R]	T
	CTC Leu	CCC Pro [P]	CAC His [H]	CGC Arg [R]	C
	CTA Leu [L]	CCA Pro [P]	CAA Gln [Q]	CGA Arg [R]	A
	CTG Leu [L]	CCG Pro [P]	CAG Gln [Q]	CGG Arg [R]	G
A	ATT Ile [I]	ACT Thr [T]	AAT Asn [N]	AGT Ser [S]	T
	ATC Ile [I]	ACC Thr [T]	AAC Asn [N]	AGC Ser [S]	C
	ATA Ile [I]	ACA Thr [T]	AAA Lys [K]	AGA Arg [R]	A
	ATG Met [M]	ACG Thr [T]	AAG Lys [K]	AGG Arg [R]	G
G	GTT Val [V]	GCT Ala [A]	GAT Asp [D]	GGT Gly [G]	T
	GTC Val [V]	GCC Ala [A]	GAC Asp [D]	GGC Gly [G]	C
	GTA Val [V]	GCA Ala [A]	GAA Glu [E]	GGA Gly [G]	A
	GTG Val [V]	GCG Ala [A]	GAG Glu [E]	GGG Gly [G]	G

(a)	Sequence 1	C	O	U	N	T	I	N	G	
	Sequence 2	N	T	I	G					
(b)	Sequence 1	C	O	U	N	T	I	N	G	Possible Alignment
	Sequence 2	-	-	-	N	T	I	G	-	(Shifting Sequence 2)
(c)	Sequence 1	C	O	U	N	T	I	N	G	Possible Alignment
	Sequence 2	-	-	-	N	T	-	I	G	(Shifting Sequence 2 and inserting a gap)
(d)	Sequence 1	C	O	U	N	T	I	N	G	Possible Alignment
	Sequence 2	-	-	-	N	T	I	-	G	(Shifting Sequence 2 and inserting a gap)

Figure 2. Possible alignments of two sequences.

search sequence databases quickly with an unknown sequence. Phylogenetic analysis is closely related to sequence alignment. The idea is to use DNA or protein sequence comparison to infer evolution history. The first step in this analysis is to perform multiple sequence alignment. Then, a phylogenetic tree is built based on the multiple alignments. The protein structure analysis involves the prediction of protein secondary and tertiary structures from the primary sequences. So far, the analyses focus on individual sequences or a handful of sequences. The next three areas are involved in system-wide analysis. Genome analysis mainly deals with the sequencing of a complete or partial genome. The problems include genome assembly, gene structure prediction, gene function annotation, and so on. Many techniques of sequence analysis are used in genome analysis, but many new methods were developed for the unique problems. Microarray technologies provide an opportunity for biologists to study the gene expression at a system level. The problems faced in the analysis are completely different from sequence analysis. Many statistical and data mining techniques are applied in the field. Network analysis is another system level study of the biological system. Biological networks can be divided into three categories: metabolic network, protein-protein interaction network, and genetic network. The questions in this area include network modeling, network inference from high throughput data, such as microarray, and network properties study. In the following several sections, the authors will provide a more in-depth discussion of each area.

SEQUENCE ALIGNMENT

Pair-Wise Sequence Alignment

Sequence alignment can be described by the following problem. Given two strings of text, X and Y (which may be DNA or amino acid sequences), find the optimal way of inserting dashes into the two sequences so as to maximize a given scoring function between them. The scoring function depends on both the length of the regions of consecutive dashes and the pairs of characters that are in the same position when gaps have been inserted. The following example from Abbas and Holmes (1) illustrates the idea of sequence alignment for two strings of text. Consider the two sequences, COUNTING and NTIG, shown in Fig. 2a. Figures 2b, 2c, and 2d show possible alignments obtained by inserting gaps (dashes) at different positions in one of the sequences. Figure 2d shows the alignment with the highest number of matching elements. The optimal alignment between two sequences depends on the scoring function that is used. As shall be shown, an optimal sequence alignment for a given scoring function may not be.

Now that what is meant by an optimal sequence alignment has been discussed, the motivation for doing so must be explained. Sequence alignment algorithms can detect mutations in the genome that lead to genetic disease and also provide a similarity score, which can be used to determine the probability that the sequences are evolutionarily related. Knowledge of evolutionary relation between a newly identified protein sequence and a family of protein sequences in a database may provide the first clues about its 3D structure and chemical function. Furthermore, by aligning families of proteins that have the same function (and may have very different sequences), a common subsequence of amino acids can be observed that is key to its particular function. These subsequences are termed protein motifs. Sequence alignment is also a first step in constructing phylogenetic trees that relate biological families of species.

A dynamic programming approach to sequence alignment was proposed by Needleman and Wunsch (2). The idea behind the dynamic programming approach can be explained using the two sequences, CCGAT and CA-AT, of Fig. 3a. If this alignment is broken into two parts (Fig. 3b),

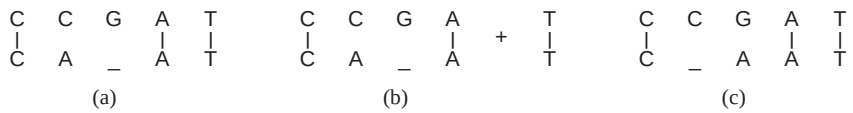


Figure 3. Overview of the dynamic programming approach.

two alignments exist: the left is the alignment of the two sequences CCGA and CA-A, and the right is the alignment of the last elements T-T. If the scoring system is additive, then the score of the alignment of Fig. 3b is the sum of the scores of the four base-alignment on the left plus the score of the alignment of the pair T-T on the right. If the alignment in Fig. 3a is optimal, then the four-base alignment in the left-hand side of Fig. 3b must also be optimal. If this were not the case (e.g., if a better alignment would be obtained by aligning A with G), then the optimal alignment of Fig. 3c would lead to a higher score than the alignment shown in Fig. 3a. The optimal alignment ending at any stage is therefore equal to the total (cumulative) score of the optimal alignment at the previous stage plus the score assigned to the aligned elements at that current stage.

The optimal alignment of two sequences ends with either the last two symbols aligned, the last symbol of one sequence aligned to a gap, or the last symbol of the other sequence aligned to a gap. In the author's analysis, x_i refers to the ith symbol in sequence 1 and y_i refers to the jth symbol in sequence 2 before any alignment has been made. The authors will use the symbol $S(i,j)$ to refer to the cumulative score of the alignment up until symbols x_i and y_j, and the symbol $s(x_i,y_j)$ to refer to the score assigned to matching elements x_i and y_j. The authors will use d to refer to the cost associated with introducing a gap.

1. If the current stage of the alignment matches two symbols, x_i and y_j, then the score, $S(i,j)$, is equal to the previous score, $S(i–1,j–1)$, plus the score assigned to aligning the two symbols, $s(x_i,y_j)$.
2. If the current match is between symbol x_i in sequence 1 and a gap in sequence 2, then the new score is equal to the score up until symbol x_{i-1} and the same symbol y_j, $S(i–1,j)$, plus the penalty associated with introducing a gap, $–d$
3. If the current match is between symbol y_j in sequence 2 and a gap in sequence 1, then the new score is equal to the previous score up until symbol y_{j-1} and the same symbol x_i, $S(i,j–1)$, plus the gap penalty $–d$

The optimal cumulative score at symbols x_i and y_j is:

$$S(i,j) = \max \begin{cases} S(i-1,j-1) + s(x_i,y_j) \\ S(i-1,j) - d \\ S(i,j-1) - d \end{cases}$$

The previous equation determines the new elements at each stage in the alignment by successive iterations from the previous stages. The maximum at any stage may not be unique. The optimal sequence alignment (s) is the one that provides the highest score, which is usually performed using a matrix representation, where the cells in the matrix are assigned an optimal score, and the optimal alignment is determined by a process called trace back (3,4).

The optimal alignment between two sequences depends on the scoring function that is used, which brings the need for a score that is biologically significant and relevant to the phenomenon being analyzed. Substitution matrices present one method of achieving this alignment using a "log-odds" scoring system. One of the first substitution matrices used to score amino acid sequences was developed by Dayhoff et al. (5). Other matrices such as the BLOSUM50 matrix (6) were also developed and use databases of more distantly related proteins.

The Needleman–Wunsch (N–W) algorithm and its variation (3) provide the best *global* alignment for two given sequences. Smith and Waterman (7) presented another dynamic programming algorithm that deals with finding the best *local* alignment for smaller subsequences of two given sequences rather than the best global alignment of the two sequences. The local alignment algorithm identifies a pair of subsegments, one from each of the given sequences, such that no other pair of subsegments exist with greater similarity.

Heuristic Alignment Methods

Heuristic search methods for sequence alignment have gained popularity and extensive use in practice because of the complexity and large number of calculations in the dynamic programming approach. Heuristic approaches search for local alignments of subsegments and use these alignments as "seeds" in which to extend out to longer sequences. The most widely used heuristic search method available today is BLAST (Basic Local Alignment Search Tool) by Altschul et al. (8). BLAST alignments define a measure of similarity called MSP (Maximal Segment Pair) as the highest scoring pair of identical length subsegments from two sequences. The lengths of the subsegments are chosen to maximize the MSP score.

Multiple Sequence Alignments

Multiple sequence alignments are alignments of more than two sequences. The inclusion of additional sequences can improve the accuracy of the alignment, find protein motifs, identify related protein sequences in a database, and predict protein secondary structure. Multiple sequence alignments are also the first step in constructing phylogenetic trees.

The most common approach for multiple alignments is progressive alignment, which involves choosing two sequences and performing a pairwise alignment of the first to the second. The third sequence is then aligned to the first and the process is repeated until all the sequences are aligned. The score of the multiple alignment is the sum of scores of the pairwise alignments. Pairwise dynamic programming can be generalized to perform multiple alignments using the progressive alignment approach; however, it is computationally impractical even when only a few sequences are involved (9). The sensitivity of progressive alignment was improved for divergent protein sequences using CLUSTAL-W (10) (available at http://clustalw.genome.ad.jp/).

Many other approaches to sequence alignment have been proposed in the literature. For example, a Bayesian

approach was suggested for adaptive sequence alignments (11,12). The data that is now available from the human genome project has suggested the need for aligning whole genome sequences where large-scale changes can be studied as opposed to single-gene insertions, deletions, and nucleotide substitutions. MuMMer (12) follows this direction and performs alignments and comparisons of very large sequences.

PHYLOGENETIC TREES

Biologists have long built trees to classify species based on morphological data. The main objectives of phylogenetic tree studies are (*1*) to reconstruct the genealogical ties between organisms and (*2*) to estimate the time of divergence between organisms since they last shared a common ancestor. With the explosion of genetic data in the last few years, tree building has become more popular, where molecular-based phylogenetic studies have been used in many applications, such as the study of gene evolution, population subdivisions, analysis of mating systems, paternity testing, environmental surveillance, and the origins of diseases that have transferred species.

From a mathematical point of view, a phylogenetic tree is a rooted binary tree with labeled leaves. A tree is binary if each vertex has either one or three neighbors. A tree is rooted if a node, R, has been selected and termed the root. A root represents an ancestral sequence from which all other nodes descend. Two important aspects of a phylogenetic tree are its topology and branch length. The topology refers to the branching pattern of the tree, and the branch length is used to represent the time between the splitting events (mutations). Figure 4a shows a rooted binary tree with six leaves. Figure 4b shows all possible distinct rooted topologies for a tree with three leaves.

The data that is used to construct trees is usually in the form of contemporary sequences and is located at the leaves. For this reason, trees are represented with all their leaves "on the ground level" rather than at different levels.

The tree-building analysis consists of two main steps. The first step, estimation, uses the data matrix to produce a tree, $\tilde{T}$, that estimates the unknown tree, T. The second step provides a confidence statement about the estimator $\tilde{T}$, which is often performed by bootstrapping methods.

Tree-building techniques can generally be classified into one of four types: distance-based methods, parsimony methods, maximum likelihood methods, and Bayesian methods. For a detailed discussion of each of these methods, see Li (13).

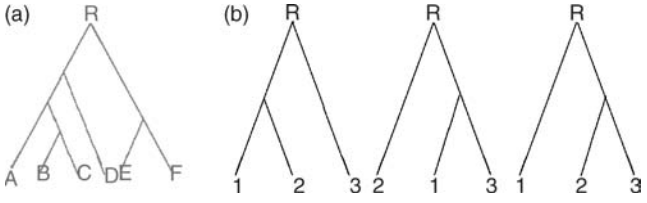


Figure 4. (a) Rooted tree with six leaves. (b) All possible topologies for three leaves.

Tree-building methods can be compared using several criteria such as accuracy (which method gives the true tree, T, when we know the answer?), consistency (when the number of characters increases to infinity, do the trees provided by the estimator converge to the true tree?), efficiency (how quickly does a method converge to the correct solution as the data size increases?), and robustness (is the method stable when the data does not fulfill the necessary assumptions?). To clarify some of these issues, read Holmes (14), where a geometric analysis of the problem is provided and these issues are further discussed.

The second part of the tree-building analysis is concerned with how close we believe the estimated tree is to the true tree. This analysis builds on a probability distribution on the space of all trees. The difficult part of this problem is that, exponentially, many possible trees exist. A nonparametric approach using a multinomial probability model on the whole set of trees would not be feasible as the number of trees is (2N-3)!!. The Bayesian approach defines parametric priors on the space of trees, and then computes the posterior distribution on the same subset of the set of all trees. This analysis enables confidence statements in a Bayesian sense (15).

PROTEIN FOLDING, SIMULATION, AND STRUCTURE PREDICTION

The main motivation for this study is that the structure of a protein greatly influences its function. Knowledge of protein structure and function can help determine the chemical structure of drugs needed to reverse the symptoms that develop due to its malfunction.

The structure of a molecule consists of atoms connected together by bonds. The bonds in a molecular structure contribute to its overall potential energy. The authors shall neglect all quantum mechanical effects in the following discussion and consider only the elements that contribute largely to the potential energy of a structure [as suggested by Levitt and Lifson (16)].

1. **Pair Bonds:** A bond that exists between atoms physically connected by a bond and separated by a distance b. It is like a spring action where energy is stored above and below an equilibrium distance, b_0. The energy associated with this bond is $U(b) = \frac{1}{2}K_b(b - b_0)^2$, where b_0 can be determined from X rays and K_b can be determined from spectroscopy.
2. **Bond Angles:** This bond exists when an angular deviation from an equilibrium angle, θ_0, occurs between three atoms. The bond angle energy associated with the triplet is $U(\theta) = \frac{1}{2}K_\theta(\theta - \theta_0)^2$.
3. **Torsion Angles:** This bond exists when a torsion angle, ϕ, exists between the first and fourth atoms on the axis of the second and third atoms. The energy associated with this bond is $U(\phi) = K_\phi(1 - \cos(n\phi + \delta))$, where θ is an initial torsion angle.
4. **Nonbonded pairs:** Bonds also exist between atoms that are not physically connected in the structure. These bonds include:

a. Van der Waal forces, which exist between nonbonded pairs and contribute to energy, $U(r) = \varepsilon[(\frac{r_0}{r})^{12} - 2(\frac{r_0}{r})^6]$, r_0 is an equilibrium distance and ε a constant.
b. Electrostatic interactions, which contribute to an energy of $U(r) = \alpha\frac{q_i q_j}{r}$; and
c. Hydrogen bonds, which result from van Der Waals forces and the geometry of the system, and contribute to the potential energy of the structure.

The total potential energy function of a given structure can thus be determined by the knowledge of the precise position of each atom. The three main techniques that are used for protein structure prediction are homology (comparative modeling), fold recognition and threading, and *ab initio* folding.

Homology or Comparative Modeling. Comparative modeling techniques predict the structure of a given protein sequence based on its alignment to one or more protein sequences of known structure in a protein database. The approach uses sequence alignment techniques to establish a correspondence between the known structure "template" and the unknown structure. Protein structures are archived for public use in an Internet-accessible database known as the Protein Data Bank (http://www.rcsb.org/pdb/) (17).

Fold Recognition and Threading. When the two sequences exhibit less similarity, the process of recognizing which folding template to use is more difficult. The first step, in this case, is to choose a structure from a library of templates in the protein databank, called fold recognition. The second step "threads" the given protein sequence into the chosen template. Several computer software programs are available for protein structure prediction using the fold recognition and threading technique such as PROSPECT (18).

Ab Initio (New Fold) Prediction. If no similarities exist with any of the sequences in the database, the *ab initio* prediction method is used. This method is one of the earliest structure prediction methods, and uses energy interaction principles to predict the protein structure (16,19,20). Some of these methods include optimization where the objective is to find a minimum energy structure (a local minimum in the energy landscape has zero forces acting on the atoms and is therefore an equilibrium state).

Monte Carlo sampling is one of the most common techniques for simulating molecular motion. The algorithm starts by choosing an initial structure, A, with potential energy, $U(A)$. A new structure, B, is then randomly generated. If the energy of the new structure is less than that of the old structure, the new structure is accepted. If the energy of the new structure is higher than the old structure, then we generate a random number, RAND, from a uniform distribution U(0,1). The new structure is accepted if $e - \frac{\Delta E}{\mathrm{KT}} > RAND$, where $\Delta E = E_B - E_A$ is the difference in energy levels, K is Boltzman's constant, and T is the temperature in kelvins. Otherwise, the new structure is rejected. Another random structure is then generated (either from the new accepted structure or from the old structure if the first one was rejected) and the process is repeated until some termination condition is satisfied (e.g., the maximal number of steps has been achieved).

Another type of analysis uses molecular dynamics uses equations of motion to trace the position of each atom during folding of the protein (21). A single structure is used as a starting point for these calculations. The force acting on each atom is the negative of the gradient of the potential energy at that position. Accelerations, a_i, are related through masses, m_i, to forces, F_i, via Netwon's second law ($F_i = m_i a_i$). At each time step, new positions and velocities of each of the atoms are determined by solving equations of motion using the old positions, old velocities, and old accelerations. Beeman (22) showed that new atomic positions and velocities could be determined by the following equations of motion

$$x(t+\Delta t) = x(t) + v(t)\Delta t + [4a(t) - a(t+\Delta t)]\frac{(\Delta t)^2}{6}$$

$$v(t+\Delta t) = v(t) + [2a(t+\Delta t) + 5a(t) - a(t-\Delta t)]\frac{\Delta t}{6}$$

where $x(t)$ = position of the atom at time t, $v(t)$ = velocity of the atom at time t, $a(t)$ = acceleration at time t, and Δt = time step in the order of 10^{-15} s for the simulation to be accurate.

In 1994, the first large-scale experiment to assess protein structure prediction methods was conducted. This experiment is known as CASP (Critical Assessment of techniques for protein Structure Prediction). The results of this experiment were published in a special issue of *Proteins* in 1995. Further experiments were developed to evaluate the fully automatic web servers for fold recognition. These experiments are known as CAFASP (Critical Assessment of Fully Automated Structure Prediction). For a discussion on the limitations, challenges, and likely future developments on the evaluation of the field of protein folding and structure prediction, the reader is referred to Bourne (23).

GENOME ANALYSIS

Analysis of completely sequenced genomes has been one of the major driving forces for the development of the bioinformatics field. The major challenges in this area include genome assembly, gene prediction, function annotation, promoter region prediction, identification of single nucleotide polymorphism (SNP), and comparative genomics of conserved regions. For a genome project, one must ask several fundamental questions: How can we put the whole genome together from many small pieces of sequences? where are the genes located on a chromosome? and what are other features we can extract from the completed genomes?

Genome Assembly

The first problem is pertaining to the genome mapping and sequence assembly. During the sequencing process, large DNA molecules with millions of base pairs, such as a

human chromosome, are broken into smaller fragments ($\sim$ 100 kb) and cloned into vector such as bacterial artificial chromosome (BAC). These BAC clones can be tiled together by physical mapping techniques. Individual BACs can be further broken down into smaller random fragments of 1–2 kb. These fragments are sequenced and assembled based on overlapping fragments. With more fragments sequenced, enough overlaps will exist to cover most of the sequence. This method is often referred as "shotgun sequencing". Computer tools were developed to assemble the small random fragments into large contigs based on the overlapping ends among the fragments using similar algorithms as the ones used in the basic sequence alignment. The widely used ones include PHRAP/Consed (24,25) and CAP3 (26). Most of the prokaryotic genomes can be sequenced directly by the shotgun sequencing strategy with special techniques for gap closure. For large genomes, such as the human genome, two strategies exist. One is to assemble large contigs first and then tile together the contigs based on the physical map to form the complete chromosome (27). Another strategy is called Whole Genome Shotgun Sequencing (WGS) strategy, which assemble the genome directly from the shotgun sequencing data in combination with mapping information (28). WGS is a faster strategy to finish a large genome, but the challenge of WGS is how to deal with the large number of repetitive sequences in a genome. Nevertheless, WGS has been successfully used in completing the *Drosophila* and human genomes (29,30).

Genome Annotation

The second problem is related to deciphering the information coded in a genome, which is often called genome annotation. The process includes the prediction of gene structures and other features on a chromosome and the function annotation of the genes. Two basic types of genes exist in a genome: RNA genes and protein encoding genes. RNA genes produce active RNA molecules such as ribosomal RNA, tRNA, and small RNA. The majority of genes in a genome are protein encoding genes. Therefore, the big challenge is how to find the protein encoding region in a genome. The simplest way to search for a protein encoding region is to search for open reading frames (ORF), which is a contiguous set of codons between two stop codons. Six possible reading frames for a given DNA sequence exist, three of which start at the first, second, and third base. The other three reading frames are at the complementary strand. The longest ORFs between the start codon and the stop codon in the same reading frame provide good, but not sufficient, evidence of a protein encoding region. Gene prediction is generally easier and more accurate in prokaryotic than eukaryotic organisms due to the intron/exon structure in eukaryote genes. Computational methods of gene prediction based on the Hidden Markov Model (HMM) have been quite successful, especially in prokaryote genome. These methods involve training a gene model to recognize genes in a particular organism. As a result of the variations in codon usage, a model must be trained for each new genome. In a prokaryote genome, genes are packed densely with relatively short intergenic sequences. The model reads through a sequence with unknown gene composition and find the regions flanked by start and stop codons. The codon composition of a gene is different from that of an intergenic region and can be used as a discriminator for gene prediction. Several software tools, such as GeneMark (31) and Glimmer (32) are widely used HMM methods in prokaryotic genome annotation. Similar ideas are also applied to eucaryote gene prediction. As a result of the intron/exon structure, the model is much more complex with more attention on the boundary of intron and exon. Programs such as GeneScan (33) and GenomeScan (34) are HMM methods for eukaryote gene prediction. Neural network-based methods have also been applied in eukaryote gene prediction, such as Grial (35). Additional information for gene prediction can be found using expressed sequence tags (ESTs), which are the sequences from cDNA libraries. As cDNA is derived from mRNA, a match to an EST is a good indication that the genomic region encodes a gene. Functional annotation of the predicted genes is another major task in genome annotation. This process can be also viewed as gene classification with different functional classification systems such as protein families, metabolic pathways, and gene ontology. The simplest way is to infer annotation from the sequence similarity to a known gene (e.g., BLAST search against a well-annotated protein database such as SWISS-PROT). A better way can be a search against protein family databases [e.g., Pfam (36)], which are built based on profile HMMs. The widely used HMM alignment tools include HMMER (37) and SAM (38). All automated annotation methods can produce mistakes. More accurate and precise annotation requires manual checking and a combination of information from different sources.

Besides the gene structures, other features such as promoters can be better analyzed with a finished genome. In prokaryotic organisms, genes involved in the same pathway are often organized in an operon structure. Finding operons in a finished genome provides information on the gene regulation. For eukaryotic organisms, the completed genomes provide upstream sequences for promoter search and prediction. Promoter prediction and detection has been a very challenging bioinformatics problem. The promoter regions are the binding sites for transcription factors (TF). Promoter prediction is to discover the sequence patterns that are specific for TF binding. Different motif finding algorithms have been applied including scoring matrix method (39), Gibbs sampling (40), and Multiple EM for Motif Elicitation (MEME) (41). The results are not quite satisfactory. Recent studies using comparative genomics methods on the problem have produced some promising results and demonstrated that the promoters are conserved among closely related species (42). In addition, microarray studies can provide additional information for promoter discoveries (see the section on microarray analysis).

Comparative Genomics

With more and more genomes being completely sequenced, comparative analysis becomes increasingly valuable and provides more insights of genome organization and

evolution. One comparative analysis is based on the orthologous genes, called clusters of orthologous groups (COG) (43). Two genes from two different organisms are considered orthologous genes if they are believed to come from a common ancestor gene. Another term, paralogous genes, refers to genes in one organism and are related to each other by gene duplication events. In COG, proteins from all completed genomes are compared. All matching proteins in all the organisms are identified and grouped into orthologous groups by speciation and gene duplication events. Related orthologous groups are then clustered to form a COG that includes both orthologs and paralogs. These clusters correspond to classes of functions. Another type of comparative analysis is based on the alignment of the genomes and studies the gene orders and chromosomal rearrangements. A set of orthologous genes that show the same gene order along the chromosomes in two closely related species is called a synteny group. The corresponding region of the chromosomes is called synteny blocks (44). In closely related species, such as mammalian species, the gene orders are highly conserved. The gene orders are changed by chromosomal rearrangements during evolution including the inversion, translocation, fusion, and fission. By comparing completely sequenced genomes, for example, human and mouse genomes, we can reveal the rearrangement events. One challenging problem is to reconstruct the ancestral genome from the multiple genome comparisons and estimate the number and types of the rearrangements (45).

MICROARRAY ANALYSIS

Microarray technologies allow biologists to monitor genome-wide patterns of gene expression in a high throughput fashion. Gene expression refers to the process of transcription. Gene expression for a particular gene can be measured as the fluctuation of the amount of messenger RNA produced from the transcription process of that gene in different conditions or samples.

DNA microarrays are typically composed of thousands of DNA sequences, called probes, fixed to a glass or silicon substrate. The DNA sequences can be long (500–1500 bp) cDNA sequences or shorter (25–70 mer) oligonucleotide sequences. The probes can be deposited with a pin or piezoelectric spray on a glass slide, known as spotted array technology. Oligonucleotide sequences can also be synthesized *in situ* on a silicon chip by photolithographic technology (i.e., Affymetrix GeneChip). Relative quantitative detection of gene expression can be carried out between two samples on one array (spotted array) or by single samples comparing multiple arrays (Affymetrix GeneChip). In spotted array experiments, samples from two sources are labeled with different fluorescent molecules (Cy3 and Cy5) and hybridized together on the same array. The relative fluorescence between each dye on each spot is then recorded and a composite image may be produced. The relative intensities of each channel represent the relative abundance of the RNA or DNA product in each of the two samples. In Affymetrix GeneChip experiments, each sample is labeled with the same dye and hybridized to different arrays. The absolute fluorescent values of each spot may then be scaled and compared with the same spot across arrays. Figure 5 gives an example of a composite image from one spotted array.

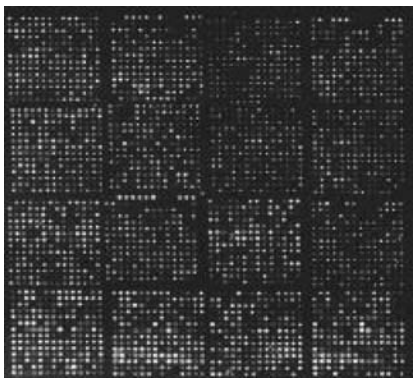


Figure 5. An image from a spotted array after laser scanning. Each spot on the image represents a gene and the intensity of a spot reflects the gene expression.

Microarray analyses usually include several steps including: image analysis and data extraction, data quantification and normalization, identification of differentially expressed genes, and knowledge discovery by data mining techniques such as clustering and classification. Image analysis and data extraction is fully automated and mainly carried out using a commercial software package or a freeware depending on the technology platforms. For example, Affymetrix developed a standard data processing procedure and software for its GeneChips (for detailed information, see http://www.affymetrix.com); GenePix is widely used image analysis software for spotted arrays. For the rest of the steps, the detailed procedures may vary depending on the experiment design and goals. We will discuss some of the procedures below.

Statistical Analysis

The purpose of normalization is to adjust for systematic variations, primarily for labeling and hybridization efficiency, so that the true biological variations can be discover as defined by the microarray experiment (46,47). For example, as shown in the self-hybridization scatter plot (Fig. 6) for a two-dye spotted array, variations (dye bias) between dyes is obvious and related to spot intensities. To correct the dye bias, one can apply the following model:

$$\log_2(R/G) \rightarrow \log_2(R/G) - c(A)$$

where R and G are the intensities of the dyes; A is the signal strength ($\log_2(R^*G)/2$); M is the logarithm ratio ($\log_2(R/G)$); $c(A)$ is the locally weighted polynomial regression (LOWESS) fit to the MA plot (48,49).

After correction of systematic variations, we want to determine which genes are significantly changed during the experiment and to assign appropriately adjusted p values to the genes. For each gene, we wish to test the null hypothesis that the gene is not differentially expressed. The P value is the probability of finding a result

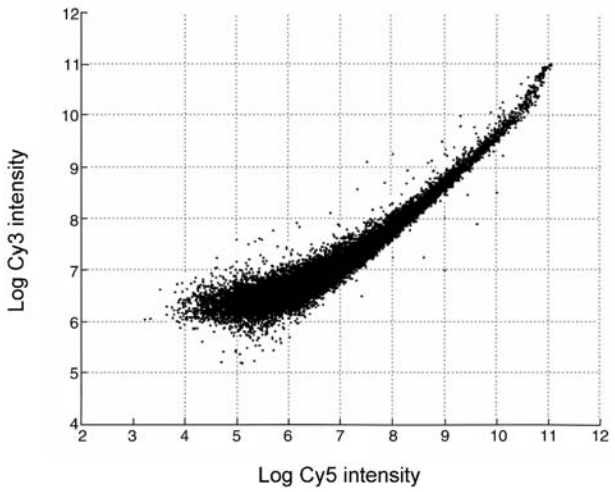


Figure 6. Self-hybridization scatter plot. The y axis is the intensity from one dye; the x axis is the intensity from the other dye. Each spot is a gene.

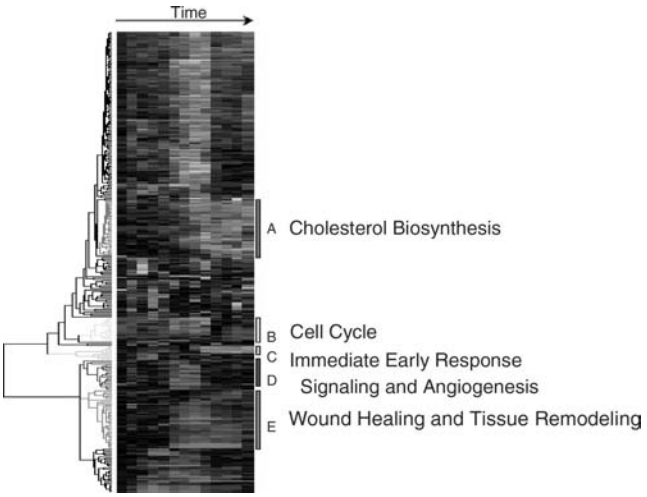


Figure 7. Hierarchical clustering of microarry data. Rows are genes. Columns are RNA samples at different time points. Values are the signals (expression levels) that are represented by the color spectrum. Green represents down-regulation whereas red represents up-regulation. The color bars beside the dendrogram show the clusters of genes that exhibit similar expression profiles (patterns). The bars are labeled with letters and description of possible biological processes involving the genes in the clusters. [Reprinted from Eisen et al. (58).]

by chance. If P value is less than a cut-off (e.g., 0.05), one would reject the null hypothesis and state that the gene is differentially expressed (50). Analysis of variance (ANOVA) is usually used to model the factors for a particular experiment. For example,

$$\log(m_{ijk}) = \mu + A_i + D_j + V_k + \varepsilon_{ijk}$$

where m_{ijk} is the ratio of intensities from the two dye-labeled samples for a gene; μ is the mean of ratios from all replicates; A is the effect of different arrays; D is the dye effects; and V is the treatment effects (51). Through F test, it will be determined if the gene exhibits differential expression between any V_k. For a typical microarray, thousands of genes exist. We need to perform thousands of tests in an experiment at the same time, which introduce the statistical problem of multiple testing and adjustment of p value. False discovery rate (FDR) (52) has been commonly adopted for this purpose.

For Affymetrix GeneChips analysis, even though the basic steps are the same as spotted microarrays, because of the difference in technology, different statistical methods were developed. Besides the statistical methods provided by Affymetrix, several popular methods are packaged into software such as dChip (53) and RMA (54) in Bioconductor (http://www.bioconductor.org). With rapid accumulation of microarray data, one challenging problem is how to compare microarray data across different technology platforms. Some recent studies on data agreements have provided some guidance (55–57).

Clustering and Classification

Once a list of significant genes is obtained from the statistical test, different data mining techniques would be applied to find interesting patterns. At this step, the microarray dataset is organized as a matrix. Each column represents a condition; each row represents a gene. An entry is the expression level of the gene under the corresponding condition. If a set of genes exhibit the similar fluctuation under all of the conditions, it may indicate that these genes are co-regulated. One way to discover the co-regulated genes is to cluster genes with similar fluctuation patterns using various clustering algorithm. Hierarchical clustering was the first clustering method applied to the problem (58). The result of hierarchical clustering forms a 2D dendrogram as shown in Fig. 7. The measurement used in the clustering process can be either a similarity, such as Pearson's correlation coefficient, or a distance, such as Euclidian distance.

Many different clustering methods have been applied later on, such as k means (59), self-organizing map (60), and support vector machine (61). Another type of microarray study involves classification techniques. For example, we can use the gene expression profile to classify cancer types. Golub et al. (62) first reported using classification techniques to classify two different types of leukemia as shown in Fig. 8. Many commercial software packages (e.g., GeneSpring and Spotfire) offer the use of these algorithms for microarray analyses.

COMPUTATIONAL MODELING AND ANALYSIS OF BIOLOGICAL NETWORKS

The biological system is a complex system involving hundreds of thousands of elements. The interaction among the elements forms an extremely complex network. With the development of high throughput technologies in functional genomics, proteomics, and metabolomics, one can start looking into the system-level mechanisms governing the interactions and properties of biological networks. Network modeling has been used extensively

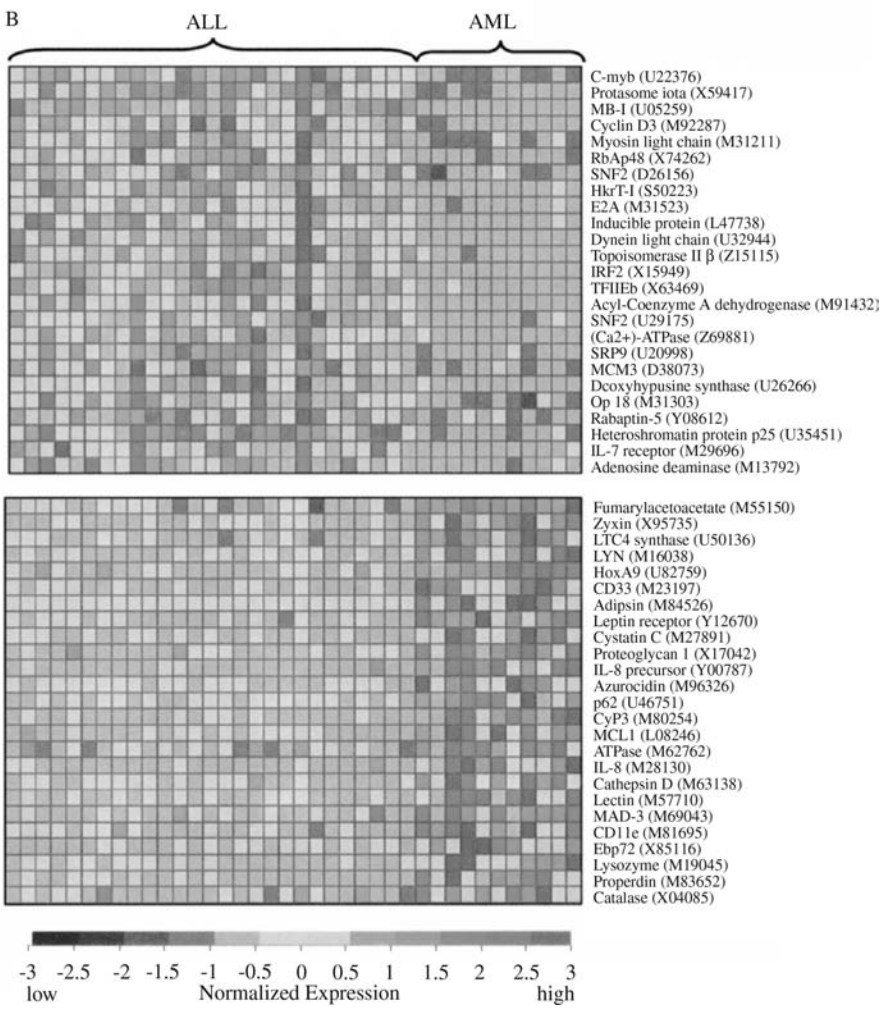


Figure 8. An example of microarray classification. Genes distinguishing acute myeloid leukemia (AML) and acute lymphoblastic leukemia (ALL). The 50 genes most highly correlated with the ALL-AML class distinction are shown. Each row corresponds to a gene, with the columns corresponding to expression levels in different samples. Expression levels for each gene are normalized across the samples such that the mean is 0 and the SD is 1. The scale indicates SDs above or below the mean. The top panel shows genes highly expressed in ALL, the bottom panel shows genes more highly expressed in AML. [Reprinted from Golub et al. (62).]

in social and economical fields for many years (63). Many methods can be applied to biological network studies.

The cellular system involves complex interactions between proteins, DNA, RNA, and smaller molecules and can be categorized in three broad subsystem, metabolic network or pathway, protein network, and genetic or gene regulatory network. *Metabolic network* represents the enzymatic processes within the cell, which provide energy and building blocks for cells. It is formed by the combination of a substrate with an enzyme in a biosynthesis or degradation reaction. Considerable information about metabolic reactions has been accumulated through many years and organized into large databases, such as KEGG (64), EcoCyc (65), and WIT (66). *Protein network* refers to the signaling networks where the basic reaction is between two proteins. Protein-protein interactions can be determined systematically using techniques such as yeast two-hybrid system (67) or derived from the text mining of literatures (68). *Genetic network or regulatory network* refers to the functional inference of direct causal gene interactions (69). One can conceptualize gene expression as a genetic feedback network. The network can be inferred from the gene expression data generated from microarray or proteomics studies in combination with computation modeling.

Metabolic network is typically represented as a graph with the vertex being all the compounds (substrates) and the edges being reactions linking the substrates. With such representation, one can study the general properties of the metabolic network. It has been shown that metabolic network exhibits typical property of small world or scale-free network (70,71). The distribution of compound connectivity follows a power law as shown in Fig. 9. Nodes serving as hubs exist in the network. Such property makes the network quite robust to random deletion of nodes, but vulnerable to selected deletion of nodes. For example, deletion of hub nodes will cause the network collapse very quickly. A recent study also shows that the metabolic network can be organized in modules based on the connectivity. The connectivity is high within modules, but low between modules (72).

Flux analysis is another important aspect in metabolic network study. Building on the stoichiometric network analysis, which only uses the well-characterized network topology, the concept of elementary flux modes was introduced (73,74). An elementary mode is a minimal set of enzymes that could operate at steady state, with the

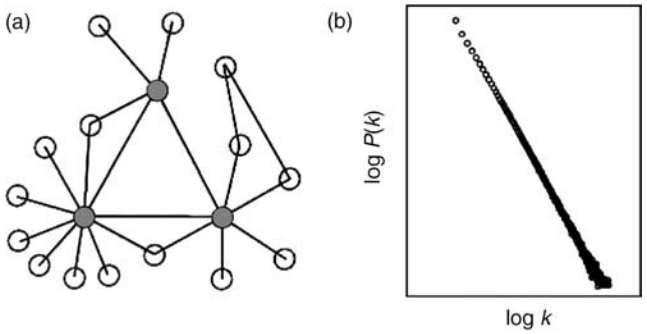


Figure 9. a. In the scale-free network, most nodes have only a few links, but a few nodes, called hubs (filled circle), have a very large number of links. b. The network connectivity can be characterized by the probability, $P(k)$, that a node has k links. $P(k)$ for a scale-free network has no well-defined peak, and for large k, it decays as a power-law, $P(k) \approx k^{-\gamma}$, appearing as a straight line with slope -γ on a log–log plot. [Reprinted from Jeong et al. (70).]

enzymes weighted by the relative flux they need to carry out the mode to function. The total number of elementary modes for given conditions has been used as a quantitative measure of network flexibility and as an estimate of fault-tolerance (75,76).

A system approach to model regulatory networks is essential to understand their dynamics. Recently, several high-level models have been proposed for the regulatory network including Boolean models, continuous systems of coupled differential equations, and probabilistic models. *Boolean networks* assume that a protein or a gene can be in one of two states, active or inactive, represented by 1 or 0. This binary state varies in time and depends on the state of the other genes and proteins in the network through a discrete equation:

$$X_i(t+1) = F_i[X_1(t), \ldots, X_N(t)] \qquad (4)$$

Thus, the function F_i is a Boolean function for the update of the ith element as a function of the state of the network at time t (69). Figure 10 gives a simple example.

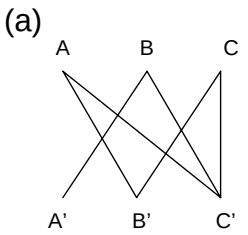
Gene expression patterns contain much of the state information of the genetic network and can be measured experimentally. We are facing the challenge of inferring or reverse engineering the internal structure of this genetic network from measurements of its output. Genes with similar temporal expression patterns may share common genetic control processes and may, therefore, be related functionally. Clustering gene expression patterns according to a similarity or distance measure is the first step toward constructing a wiring diagram for a genetic network (78).

Differential equations can be an alternative model to the Boolean network and applied when the state variables X are continuous and satisfy a system of differential equations of the form

$$\frac{dX_i}{dt} = F_i[X_1(t), \ldots, X_N(t), I(t)]$$

where the vector *I(t)* represents some external input into the system. The variable X_i can be interpreted as representing concentrations of proteins or mRNAs. Such a model has been used to model biochemical reactions in the metabolic pathways and gene regulation (69).

Bayesian networks are provided by the theory of graphical models in statistics. The basic idea is to approximate a complex multidimensional probability distribution using a product of simpler local probability distributions. Generally, a Bayesian network model is based on a directed acyclic graph (DAG) with N nodes. In genetic network, the nodes may represent genes or proteins and the random variables X_i levels of activity. The parameters of the model are the local conditional distributions of each random variable given the random variables associated with the



(b)

A'= B
B'= A OR C
C'= (A AND B) OR (B AND C) OR (A AND C)

(c)

Input			Output		
A	B	C	A'	B'	C'
0	0	0	0	0	0
0	0	1	0	1	0
0	1	0	1	0	0
0	1	1	1	1	1
1	0	0	0	1	0
1	0	1	0	1	1

Figure 10. Target Boolean network for reverse engineering. (a) The network wiring and (b) logical rules determine (c) the dynamic output. The challenge lies in inferring (a) and (b) from (c). [Reprinted from Liang et al. (77).]

parent nodes

$$P(X_1, \ldots, X_N) = \prod_i P(X_i | X_j : j \in N^-(i)) \quad (4)$$

where $N^-(i)$ denotes all the parents of vertex i. Given a dataset D representing expression levels derived using DNA microarray experiments; it is possible to use learning techniques with heuristic approximation methods to infer the network architecture and parameters. As data from microarray experiments are still limited and insufficient to completely determine a single model, people have developed heuristics for learning classes of models rather than single models, for instance, for a set of co-regulated genes (69). Bayesian networks have recently been shown to combine heterogeneous datasets, for instance, microarray data with functional annotation and mutation data to produce an expert system (79).

In this chapter, some major development in the field of bioinformatics were reviewed and some basic concepts in the field were introduced covering six areas: sequence analysis, phylogenetic analysis, protein structure analysis, genome analysis, microarray analysis, and network analysis. Due to the limited space, some topics have been left out. One such topics is text mining, which uses Natural Language Processing (NLP) techniques to extract information from the vast amount of literature in biological research. Text mining has become an integral part in bioinformatics. With the continuing development and maturing of new technologies in many system-level studies, the way that biological research is conducted is undergoing revolutionary change. Systems biology is becoming a major theme and driving force. The challenges for bioinformatics in the post-genomics era lie on the integration of data and knowledge from heterogeneous sources and system-level modeling and simulation providing molecular mechanism for physiological phenomena.

BIBLIOGRAPHY

Cited References

1. Abbas A, Holmes S. Bioinformatics and management science. Some common tools and techniques. Operations Res 2004; 52(2):165–190.
2. Needleman SB, Wunsch CD. A general method applicable to the search for similarities in amino acid sequence of two proteins. J Mol Biol 1970;48:443–453.
3. Gotoh O. An improved algorithm for matching biological sequences. J Mol Biol 1982;162:705–708.
4. Durbin S, Eddy S, Krogh A, Mitchison G. Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids. Cambridge, (UK): Cambridge University Press; 1998.
5. Dayhoff MO, Schwartz RM, Orcutt BC. A model of evolutionary change in proteins. Atlas of Protein Sequence and Structure. vol. 5, supplement 3. National Biomedical Research Foundation. Washington, (DC): 1978. p 345–352.
6. Henikoff S, Henikoff JG. Amino acid substitution matrices from protein blocks. Proc Natl Acad Sci USA 1992;89:10915–10919.
7. Smith TF, Waterman MS. Identification of common molecular subsequences. J Mol Biol 1981;147:195–197.
8. Altschul SF, Gish W, Miller W, Myers E, Lipman J. Basic local alignment search tool. J Molec Biol 1990;215:403–410.
9. Lipman JD, Altschul SF, Kececioglu JD. A tool for multiple sequence alignment. Proc Natl Acad Sci 1989;86:4412–4415.
10. Thompson JD, Higgins DG, Gibson TJ. CLUSTAL W: Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position specific gap penalties and weight matrix choice. Nucleic Acids Res 1994;22: 4673–4680.
11. Lawrence CE, Altschul SF, Boguski MS, Liu JS, Neuwald AN, Wootton J. Detecting subtle sequence signals: A Gibbs sampling strategy for multiple alignment. Science 1993;262:208–214.
12. (a) Delcher et al. 2002. (b) Zhu J, Liu JS, Lawrence CE. Bayesian adaptive sequence alignment algorithms. Bioinformatics 1998; 14:25–39.
13. Li WH. Molecular Evolution. Boston, MA: Sinauer Associates; 1997.
14. Holmes S. Bootstrapping phylogenetic trees. To appear in Statistical Science. Submitted in (2002).
15. Li S, Pearl DK, Doss H. Phylogenetic tree construction using MCMC. J Am Statist Assoc 2000;95:493–503.
16. Levitt M, Lifson S. Refinement of protein confirmations using a macromolecular energy minimization procedure. J Mol Biol 1969;46:269–279.
17. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. The protein data bank. Nucleic Acids Res 2000;28:235–242.
18. Xu Y, Xu D. Protein threading using PROSPECT: Design and evaluation. Proteins Structure, Function, and Genetics 2000;40:343–354.
19. Levitt M, Warshel A. Computer simulation of protein folding. Nature 1975;253:694–698.
20. Nemethy G, Scheraga HA. Theoretical determination of sterically allowed conformations of a polypeptide chain by a computer method. Biopolymers 1965;3:155–184.
21. Levitt M. Molecular dynamics of native protein: Computer simulation of the trajectories. J Mol Biol 1983;168:595–620.
22. Beeman D. Some multi-step methods for use in molecular dynamics calculations. J Comput Phys 1976;20:130–139.
23. Bourne PE. CASP and CAFASP experiments and their findings. Methods Biochem Anal 2003;44:501–507.
24. Gordon D, Abajian C, Green P. Consed: A graphical tool for sequence finishing. Genome Res 1998;8(3):195–202.
25. Gordon D, Desmarais C, Green P. Automated finishing with autofinish. Genome Res 2001;11(4):614–625.
26. Huang X, Madan A. CAP3: A DNA sequence assembly program. Genome Res 1999;9(9):868–877.
27. Waterston RH, Lander ES, Sulston JE. On the sequencing of the human genome. Proc Natl Acad Sci USA 2002; 99(6):3712–3716.
28. Myers EW, et al. A whole-genome assembly of Drosophila 2000;287(5461):2196–2204.
29. Adams MD, et al. The genome sequence of Drosophila melanogaster Science 2000;287(5461):2185–2195.
30. Venter JC, et al. The sequence of the human genome. Science 2001;29:1304–1351.
31. Lukashin AV, Borodovsky M. GeneMark.hmm: New solutions for gene finding. Nucleic Acids Res 1998;26(4):1107–1115.
32. Delcher AL, Harmon D, Kasif S, White O, Salzberg SL. Improved microbial gene identification with GLIMMER. Nucleic Acids Res 1999;27(23):4636–4641.
33. Burge C, Karlin S. Prediction of complete gene structures in human genomic DNA. J Mol Biol 1997;268:78–94.
34. Yeh R-F, Lim LP, Burge CB. Computational inference of homologous gene structures in the human genome. Genome Res 2001;11:803–816.

35. Xu Y, Uberbacher CE. Automated gene identification in large-scale genomic sequences. J Comp Biol 1997;4:325–338.
36. Bateman A, Coin L, Durbin R, Finn RD, Hollich V, Griffiths-Jones S, Khanna A, Marshall M, Moxon S, Sonnhammer ELL, Studholme DJ, Yeats C, Eddy SR. The Pfam Protein Families Database. Nucleic Acids Res 2004;32:D138–D141.
37. Eddy S. Profile hidden Markov models. Bioinformatics 1998;14:755–763.
38. Krogh A, Brown M, Mian IS, Juolander K, Haussler D. Hidden Markov models in computational biology applications to protein modeling. J Mol Biol 1994;235: 1501–1531.
39. Stomo GD, Hartzell GW. Identifying protein-binding sites from unaligned DNA fragments. Proc Natl Acad Sci 1989;86:1183–1187.
40. Lawrence CE, Reilly AA. An expectation maximization (EM) algorithm for the identification and characterization of common sites in unaligned biopolymer sequences. Proteins Struct Funct Genet 1990;7:41–51.
41. Bailey LT, Elkan C. Fitting a mixture model by expectation maximization to discover motifs in biopolymers. Proceedings of the Second International Conference on Intelligent Systems for Molecular Biology 1994; 28–36.
42. Kellis M, Birren BW, Lander ES. Proof and evolutionary analysis of ancient genome duplication in the yeast Saccharomyces cerevisiae. Nature 2004;428:617–624.
43. Tatusov RL, Koonin EV, Lipman DJ. A genomic perspective on protein families. Science 1997; 631–637.
44. O'Brien SJ, Menotti-Raymond M, Murphy WJ, Nash WG, Wienberg J, Stanyon R, Copeland NG, Jenkins NA, Womack JE, Graves JAM. The promise of comparative genomics in mammals. Science 1999;286:458–481.
45. Bourque G, Pevzner AP. Genome-scale evolution: Reconstructing gene orders in the ancestral species. Genome Res 2002;12:26–36.
46. Bolstad BM, Irizarry RA, Astrand M, Speed TP. A comparison of normalization methods for high density oligonucleotide array data based on bias and variance. Bioinformatics 2003;19(2):185–193.
47. Bajesy et al. 2005.
48. Yang YH, Dudoit S, Luu P, Lin DM, Peng V, Ngai J, Speed TP. Normalization for cDNA microarray data: A robust composite method addressing single and multiple slide systematic variation. Nucleic Acids Res 2002;30(4):e15.
49. Yang YH, Thorne N. Normalization for two-color cDNA microarray data. Science and statistics: A festschrift for terry speed. In: Goldstein D, ed. IMS Lecture Notes, Monograph Series. Vol. 40; 2003. p 403–418.
50. Smyth GK, Yang YH, Speed TP. Statistical issues in microarray data analysis. In: Brownstein MJ, Khodursky AB, eds. Functional Genomics: Methods and Protocols. Methods in Molecular Biology. vol. 224. Totowa, (NJ): Humana Press; 2003. p 111–136.
51. Kerr M, Churchill G. Analysis of variance for gene expression microarray data. J Comp Biol 2000;7:819–837.
52. Benjamini Y, Hochberg Y. Controlling the false discovery rate: A practical and powerful approach to multiple testing. J Royal Statist Soc B 1995;57(1):289–300.
53. Li C, Wong WH. Model-based analysis of oligonucleotide arrays: Expression index computation and outlier detection. Proc Natl Acad Sci 2001;98:31–36.
54. Bolstad BM, Irizarry RA, Astrand M, Speed TP. A comparison of normalization methods for high density oligonucleotide array data based on bias and variance. Bioinformatics 2003;19(2):185–193.
55. Wang H, He X, Band M, Wilson C, Liu L. A study of inter-lab and inter-platform agreement of DNA microarray data. BMC Genomics 2005;6(1):71.
56. Jarvinen A, Hautaniemi S, Edgren H, Auvinen P, Saarela J, Kallioniemi O, Monni O. Are data from different gene expression microarray platforms comparable? Genomics 2004;83: 1164–1168.
57. Culhane AC, Perriere G, Higgins DG. Cross-platform comparison and visualisation of gene expression data using coinertia analysis. BMC Bioinformatics 2003;4:59.
58. Eisen MB, Spellman PT, Brown PO, Botstein D. Cluster analysis and display of genome-wide expression patterns. Proc Natl Acad Sci USA 1998;95(25):14863–14868.
59. Ben-Dor A, Shamir R, Yakhini Z. Clustering gene expression patterns. J Comp Biol 1999;6(3/4):281–297.
60. Tamayo P, Solni D, Mesirov J, Zhu Q, Kitareewan S, Dmitrovsky E, Lander ES, Golub TR. Interpreting patterns of gene expression with self-organizing maps: Methods and application to hematopoietic differentiation. Proc Natl Acad Sci USA 1999;96(6):2907–2912.
61. Alter O, Brown PO, Bostein D. Singular value decomposition for genome-wide expression data processing and modeling. Proc Natl Acad Sci USA 2000;97(18):10101–10106.
62. Golub TR, Slonim DK, Tamayo P, Huard C, Gaasenbeek M, Mesirov JP, Coller H, Loh ML, Downing JR, Caligiuri MA, Bloomfield CD, Lander ES. Molecular classification of cancer: Class discovery and cass prediction by gene expression monitoring. Science 1999;286:531–537.
63. Sole RV, Ferrer-Cancho R, Montoya JM, Valverde S. Selection, tinkering, and emergence in complex networks. Complexity 2003;8:20–33.
64. Kanehisa M. A database for post-genome analysis. Trends Genet 1997;13:375–376.
65. Keseler IM, Collado-Vides J, Gama-Castro S, Ingraham J, Paley S, Paulsen IT, Peralta-Gil M, Karp PD. EcoCyc: A comprehensive database resource for Escherichia coli. Nucleic Acids Res 2005;33:D334–D337.
66. Overbeek R, Larsen N, Pusch GD, D'Souza M, Selkov E, Kyrpides N, Fonstein M, Maltsev N, Selkov E. WIT: Integrated system for high-throughput genome sequence analysis and metabolic reconstruction. Nucleic Acids Res 2000;28(1): 123–125.
67. Fields S, Song OK. A novel genetic system to detect protein-protein interactions. Nature 1989;340:245–246.
68. Daraselia N, Yuryev A, Egorov S, Novichkova S, Nikitin A, Mazo I. Extraction of human protein interactions from MEDLINE using full-sentence parser. Bioinformatics 2003;19:1–8.
69. Baldi P, Hatfield GW. Microarrays and Gene Expression. Cambridge, (UK): Cambridge University Press; 2001.
70. Jeong H, Tombor B, Albert1 R, Oltvai ZN, Barabási AL. The large-scale organization of metabolic networks. Nature 2000; 407:651–654.
71. Wagner A, Fell DA. The small world inside large metabolic networks. Proc Royal Soc Lond B 2001;268:1803–1810.
72. Guimera R, Nunes Ameral AL. Functional cartography of complex metabolic networks. Nature 2005;433:895–900.
73. Schuster S, Hilgetag C, Woods JH, Fell DA. Reaction routes in biochemical reaction systems: algebraic properties, validated calculation procedure and example from nucleotide metabolism. J Math Biol 2002;45(2):153–181.
74. Schuster S, Fell DA, Dandekar T. A general definition of metabolic pathways useful for systematic organization and analysis of complex metabolic networks. Nature 2000;18: 326–332.
75. Stelling J, Klamt S, Bettenbrock K, Schuster S, Gilles ED. Metabolic network structure determines key aspects of functionality and regulation. Nature 2002;420:190–193.

76. Cakir T, Kirdar B, Ulgen KO. Metabolic pathway analysis of yeast strengthens the bridge between transcriptomics and metabolic networks. Biotechnol Bioeng 2004;86:251–260.
77. Liang S, Fuhrman S, Somogyi R. REVEAL, a general reverse engineering algorithm for inference of genetic network architectures. Pacific Symp Biocomput 1998;3:18–29.
78. Somogyi R, Fuhrman S, Wen X. Genetic network inference in computational models and applications to large-scale gene expression data. Cambridge, (MA): MIT Press; 2001.
79. Troyanskaya OG, Dolinski K, Owen AB, Altman RB, Botstein D. A Bayesian framework for combining heterogeneous data sources for gene function prediction (in Saccharomyces cerevisiae). Proc Natl Acad Sci USA 2003;100: 8348–8353.

Further Reading

Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. Nucleic Acids Res 1997;25:3389–3402.

Baldi P, Chauvin Y, Hunkapillar T, McClure M. Hidden Markov models of biological primary sequence information. Proc Natl Acad Sci USA 1994;91:1059–1063.

Baldi P, Brunak S. Bioinformatics: The Machine Learning Approach. 2nd ed. Cambridge, (MA): MIT Press; 2001.

Bork P, Dandekar T, Diaz-Lazcoz Y, Eisenhaber F, Huynen M, Yuan Y. Predicting function: From genes to genomes and back. J Mol Biol 1998;283:707–725.

Bower J, Bolouri H. Computational Modeling of Genetic and Biochemical Networks. Cambridge, (MA): MIT Press; 2001.

Bray N, Dubchak I, Pachter L. AVID: A global alignment program. Genome Res 2003;13(1):97–102.

Brown PO, Botstein D. Exploring the new world of the genome with DNA microarrays. Nature Genetics 1999;21:33–37.

Brudno M, CB Do, Cooper GM, Kim MF, Davydov E, Green ED, Sidow A, Batzoglou A. LAGAN and Multi-LAGAN: efficient tools for large-scale multiple alignment of genomic DNA. Genome Res 2003;13:(4):721–731.

Brudno M, Malde S, Poiakov A, Do C, Couronne O, Dubchak I, Batzoglou A. Glocal alignment: Finding rearrangements during alignment. Bioinformatics Special Issue on the Proceedings of the ISMB 2003;19:54i–62i.

Bryant SH, Altschul SF. Statistics of sequence-structure threading. Curr Opin Structur Biol 1995;5:236–244.

Cohen FE. Protein misfolding and prion diseases. J Mol Biol 1999;293:313–320.

Diaconis P, Holmes S. Random walks on trees and matchings. Electron J Probabil 2002;7:1–17.

Doyle JC. Robustness and dynamics in biological networks. In: The First International Conference on Systems Biology. New York: Japan Science and Technology Corporation, MIT Press; 2000.

Dudoit S, Fridlyand J, Speed TP. Comparison of discrimination methods for the classification of tumors using gene expression data. J Am Statistic Assoc 2002;97:77–87.

Eddy S, Mitchison G, Durbin R. Maximum discrimination hidden Markov models of sequence consensus. J Comput Biol 1995; 2:9–23.

Eddy SR. Non-coding RNA genes and the modern RNA world. Nature Rev Genet 2001;2:919–929.

Efron B, Halloran EE, Holmes S. Bootstrap confidence levels for phylogenetic trees. Proc Natl Acad Sci 1996;93:13429–13434.

Farris JS. The logical basis of phylogenetic analysis. In: Platnick N, Funk V, eds. Advances in Cladistics. vol. 2. 1983. p 7–36.

Fedorov AN, Baldwin TO. Contranslational protein folding. Biol Chem 1997;272(52):32715–32718.

Felsenstein J. Evolutionary trees from DNA sequences: A maximum likelihood approach. J Mol Evol 1981;17(6):368–376.

Felsenstein J. 1993. (Phylogeny Inference Package) version 3.5c. Department of Genetics, University of Washington, Seattle, WA. Available http://evolution.genetics.washington.edu/phylip.html.

Fischer D, Barret C, Bryson K, Elofsson A, Godzik A, Jones D, Karplus KJ, Kelley LA, MacCallum RM, Pawowski K, Rost B, Rychlewski L, Sternberg M. CAFASP-1: Critical assessment of fully automated structure prediction methods. Proteins 1999;3:209–217.

Fitch WM, Margoliash E. Construction of phylogenetic trees. Science 1967;155:279–284.

Foulds LR, Graham RL. The Steiner problem in Phylogeny is NP-complete. Adv Appl Math 1982;3:43–49.

Friedman N, Linial M, Nachman I, Peter D. Using Bayesian networks to analyze expression data. J Comp Bio 2000;7: 601–620.

Gardner M. The Last Recreations. New York: Copernicus-Springer Verlag; 1997.

Geman S, Geman D. Stochastic relaxation, Gibbs distribution and the Bayesian restoration of images. IEEE Trans Pattern Anal Machine Intell 1984;6:721–741.

Gibson KD, Scheraga HA. Revised algorithms for the build-up procedure for predicting protein conformations by energy minimization. J Comp Chem 1987;9:327–355.

Goloboff PA. SPA. 1995. (S)ankoff (P)arsimony (A)nalysis, version 1.1. Computer program distributed by J. M. Carpenter, Department of Entomology, American Museum of Natural History, New York.

Gribaldo S, Cammarano P. The root of the universal tree of life inferred from anciently duplicated genes encoding components of the protein-targeting machinery. J Mol Evol 1998;47(5):508–516.

Haeckel E. Morphologie der Organismen: Allgemeine Grundzuge der Organischen FormenWissenschaft, Mechanisch Begrundet durch die von Charles Darwin Reformirte Descendenz-Theorie. Berlin: Georg Riemer; 1866.

Hannenhalli S, Pevzner PA. Transforming cabbage into turnip: Polynomial algorithm for sorting signed permutations by reversals. STOC 1995; 178–189.

Helden JV, Andre B, Collado-Vides J. Extracting regulatory sites from the upstream region of yeast genes by computational analysis of oligonucleotide frequencies. J Mol Bio 1998;281: 827–842.

Hooper E. The River. Boston, (MA): Little, Brown; 1999.

Huelsenbeck J, Ronquist F. 2002. Mr. Bayes. Bayesian inference of phylogeny. Available at http://morphbank.ebc.uu.se/mrbayes/links.php.

Jukes T, Cantor C. Evolution of protein molecules. In: eds. Munro HN, Mammalian Protein Metabolism. New York: Academic Press; 1969. p 21–132.

Karlin S, Altschul SF. Methods for assessing the statistical significance of molecular sequences features by using general scoring schemes. Proc Natl Acad Sci USA 1990;87(6):2264–2268.

Keith JM, Adams P, Bryant D, Kroese DP, Mitchelson KR, Cochran DAE, Lala GH. A simulated annealing algorithm for finding consensus sequences. Bioinformatics 2002;18: 1494–1499.

Kent WJ. BLAT–the BLAST-like alignment tool. Genome Res 2002;12(4):656–664.

Kirkpatrick S, Gelatt CD Jr, Vecchi MP. Optimization by simulated annealing. Science 1983;220:671–680.

Korf I, Flicek P, Duan D, Brent MR. Integrating genomic homology into gene structure prediction. Bioinformatics 2001;17:S140–S148.

Levitt M. Protein folding by restrained energy minimization and molecular dynamics. J Mol Biol 1983;170:723–764.
Ly DH, Lockhart DJ, Lerner RA, Schultz PG. Mitotic misregulation and human aging. Science 2000;287:1241–1248.
Ma B, Tromp J, Li M. PatternHunter: Faster and more sensitive homology search. Bioinformatics 2002;18:440–445.
Ma B, Wang Z, Zhang K. Alignment between two multiple alignments. In: Combinatorial Pattern Matching: 14th Annual Symposium, CPM 2003, Morelia, Michoacán, Mexico, June 25–27. Lecture Notes in Computer Science, vol. 2676. Heidelberg, Germany: Springer-Verlag; 2003.
Maddison D, Maddison W. 2002. Sinauer. Available at http://phylogeny.arizona.edu/macclade.
McAdams H, Shapiro L. Circuit simulation of genetic networks. Science 1995;269:650–656.
Metropolis N, Rosenbluth A, Rosenbluth M, Teller A, Teller E. Simulated Annealing. J Chem Phys 1953;21:1087–1092.
Mjolsness E, Sharp DH, Rinetz J. A connectionsit model of development. J Theor Biol 1991;152:429–453.
Morales LB, Garduno-Juarez R, Romero D. Applications of simulated annealing to the multiple-minima problem in small peptides. J Biomol Struc Dyn 1991;8:721–735.
Morgenstern B. Dialign2: improvement of the segment-to-segment approach to multiple sequence alignment. Bioinformatics 1999;15:211–218.
Mountain JL, Cavalli-Sforza LL. Inference of human evolution through cladistic analysis of nuclear DNA restriction polymorphisms. Proc Natl Acad Sci USA 1994;91:6515–6519.
Muckstein U, Hofacker IL, Stadler PF. Stochastic pairwise alignments. Bioinformatics 2002;18(sup. 2):S153–S160.
Notredame C, Higgins D, Heringa J. T-Coffee: A novel method for multiple sequence alignments. J Mol Biol 2000;302:205–217.
Peitsch MC. ProMod and Swiss-Model: Internet-based tools for automated comparative protein modeling. Biochem Soc Trans 1996;24:274–279.
Pevzner PA. Computational Molecular Biology, an Algorithmic Approach. Cambridge, (MA): MIT Press; 2000.
Pieper U, Eswar N, Ilyin VA, Stuart A, Sali A. ModBase, a database of annotated comparative protein structure models. Nucleic Acids Res 2002;30:255–259.
Ramachandran GN, Sasisekharan V. Conformation of polypeptides and proteins. Adv Protein Chem 1968;23:283–438.
Rannala B, Yang Z. Probability distribution of molecular evolutionary trees: A new method of phylogenetic inference. J Mol Evol 1996;43:304–311.
Richards FM. The protein folding problem. Sci. Am 1991; January: 54–63.
Saitou N, Nei M. The neighbor-joining method: A new method for reconstructing phylogenetic trees. Mol Biol Evol 1987;4(4): 406–425.
Schlick T. Optimization methods in computational chemistry. In: Reviews in Computational Chemistry, III. New York: VCH Publishers; 1992. p 1–71.
Schmulevich I, Dougherty E, Kim S, Zhang W. Probabilistic Boolean networks: A rule-based uncertainty model for gene regulatory networks. Bioinformatics 2002;18:261–274.
Schröder E. Vier combinatorische probleme. Z Math Phys 1870;15:361–376.
Shannon CE. A mathematical theory of communication. Bell Sys Tech J 1948;27:379–423, 623–656.
Snow ME. Powerful simulated annealing algorithm locates global minima of protein folding potentials from multiple starting conformations. J Comput Chem 1992;13:579–584.
Stanley R. Enumerative Combinatorics. vol. I. 2nd ed. Cambridge (MA): Cambridge University Press; 1996.
Swofford DL. PAUP. Phylogenetic analysis using parsimony. V4.0. Boston, (MA): Sinauer Associates; 2001.
Tozeren A, Byers SW. New Biology for Engineers and Computer Scientists. Englewood Cliffs, (NJ): Prentice Hall; 2003.
Wang LS, Jansen R, Moret B, Raubeson L, Warnow T. Fast phylogenetic methods for the analysis of genome rearrangement data: An empirical study. Proc of 7th Pacific Symposium on Biocomputing, 2002.
Watson JD, Crick FH. A structure for deoxyribose nucleic acid. Nature 1953; April.
White KP, Rifkin SA, Hurban P, Hogness DD. Microanalysis of drosphila development during metamorphosis. Science 1999; 286:2179–2184.
Winkler H. Verbeitung und Ursache der Parthenogenesis im Pflanzen und Tierreiche. Jena: Verlag Fischer; 1920.
Xu J, Hagler A. Review: Chemoinformatics and drug discovery. Molecules 2002;7:566–600.
Yang Z, Rannala B. Bayesian phylogenetic inference using DNA sequences: A Markov chain Monte Carlo method. Mol Biol Evol 1997;14:717–724.

See also COMPUTERS IN THE BIOMEDICAL LABORATORY; DNA SEQUENCE; MEDICAL EDUCATION, COMPUTERS IN; POLYMERASE CHAIN REACTION; STATISTICAL METHODS.

BIOLOGIC THERAPY.

See IMMUNOTHERAPY.

BIOMAGNETISM

DOUGLAS CHEYNE
Hospital for Sick Children
Research Institute

JINI VRBA
VSM MedTech Ltd.

INTRODUCTION

The science of *biomagnetism* refers to the measurement of magnetic fields produced by living organisms. These tiny magnetic fields are produced by naturally occurring electric currents resulting from muscle contraction, or signal transmission in the nervous system, or by the magnetization of biological tissue. The first observation of biomagnetic activity in humans was the recording of the magnetic field produced by the electrical activity of the heart, or *magnetocardiogram*, by Baule and McFee in 1963 (1). In 1968, David Cohen (2) at the Massachusetts Institute of Technology reported the first measurement of the alpha rhythm of the human brain, demonstrating that it was possible to measure magnetic fields of biological origin that are only several hundred femtotesla in magnitude (1 femtotesla $= 10^{-15}$ T)—more than 1 million times smaller than the earth's magnetic field ($\sim 5 \times 10^{-5}$ T). These early measurements were achieved using crude instruments consisting of inductance coils of 1–2 million windings in magnetically shielded enclosures and using extensive signal averaging. Instruments with increased sensitivity and performance based on the *superconducting quantum interference device*, or SQUID became available shortly after these pioneering measurements. The SQUID is a highly sensitive magnetic flux detector based on the

properties of electrical currents flowing in superconducting circuits, as predicted by Nobel laureate Brian Josephson in 1962 (3). The SQUID was soon adapted for use in biomagnetic measurements (4) and by the early 1970s, measurements of the spontaneous activity of the human heart (5) and brain (6) had been achieved without the need for signal averaging using superconducting sensing coils coupled to SQUIDs immersed in cryogenic vessels containing liquid helium. Thereafter, the field of biomagnetism continued to expand with the further development of SQUID based instrumentation during the 1970s and 1980s. The introduction in 1992 of multichannel biomagnetometers capable of simultaneous measurement of neuromagnetic activity from the entire the human brain (7,8) has resulted in widespread interest in the field of *magnetoencephalography* or *MEG* as a new method of studying human brain function.

Biomagnetic measurements are considered to have a number of advantages over more traditional electrophysiological measurements of heart and brain activity, such as the electrocardiogram or electroencephalogram. One significant advantage is that propagation of magnetic fields through the body is less distorted by the varying conductivities of the overlying tissues in comparison to electrical potentials measured from the surface of the scalp or torso, and can therefore provide a more precise localization of the underlying generators of these signals. In applications such as MEG and magnetocardiography (MCG), these measurements are completely passive and can be made repeatedly without posing any risk or harm to the patient. Also, biomagnetic signals are a more direct measure of the underlying currents in comparison to surface electrical recordings that measure volume conducted activity that must be subtracted from a reference potential at another location complicating the interpretation of the signal. In addition, magnetic measurements from multiple sites can be less time consuming since there is no need to affix electrodes to the surface of the body. As a result, biomagnetic measurements provide an accurate and noninvasive method for locating sources of electrical activity in the human body. The development of multichannel MEG systems has dramatically increased the usefulness of this technology in clinical assessment and treatment of various brain disorders. This has resulted in the recognition of routine clinical procedures by health agencies in the United States for the use of MEG to map sensory areas of the brain or localize the origins of seizure activity prior to surgery. Clinical applications of MCG have also been developed although to a lesser extent than MEG. This includes the assessment of coronary artery disease and other disorders affecting the propagation of electrical signals in the human heart. Another biomagnetic technique, known as *biosusceptometry*, involves measuring magnetized materials in the human body by measuring their moment as they are moved within a strong magnetic field. These measures can provide useful information regarding the concentration of ferromagnetic or strongly paramagnetic materials in various organs of the body, such as iron particles in the lung or iron-containing proteins in the liver. In addition, novel biomagnetometer systems are now available for the assessment of fetal brain and heart function in utero, and may provide a new clinical tool for the assessment of fetal health. Currently, there are >100 multichannel MEG systems worldwide and advanced magnetometer systems specialized for the measurement of magnetic signals from the heart, liver, lung, peripheral nervous system, as well as the fetal heart and fetal brain are currently being commercially developed. Although biomagnetism is still regarded as a relatively new field of science, new applications of biomagnetic measurements in basic research and clinical medicine are rapidly being developed, and may provide novel methods for the assessment and treatment of a variety of biological disorders. The following section reviews the current state of biomagnetic instrumentation and signal processing and its application to the measurement of human biological function.

BIOMAGNETIC INSTRUMENTATION

SQUID Sensors and Electronics

The SQUID sensor is the heart of a biomagnetometer system and provides high sensitivity detection of very small magnetic signals. The most popular types of SQUIDs are direct current (dc) and radio frequency (rf) SQUIDs, deriving their names from the method of their biasing. The modern commercial biomagnetometer instrumentation uses dc SQUIDs implemented in low temperature superconducting materials (usually Nb). In recent years, there has been significant progress in the development of high Tc SQUIDs, both dc and rf. These devices are usually constructed from $YBa_2Cu_3O_{7-x}$ ceramics. However, due to their poorer low frequency performance and difficulties with reproducible large volume manufacturing they are not yet suitable for large-scale applications. An excellent review of SQUID operation can be found in (9).

The rf SQUID was popular in the early days of superconducting magnetometry because they required only one Josephson junction. However, in majority of low Tc commercial applications, the rf SQUIDs have been displaced by dc SQUIDs due to their greater sensitivity, although in recent years, interest in rf SQUIDs has been renewed in connection with high Tc superconductivity. The operation of SQUIDs is illustrated in Fig. 1a. The dc SQUID can be modeled as a superconducting ring interrupted by two resistively shunted Josephson junctions as in Fig. 1a (11). The Josephson junctions are superconducting quantum mechanical devices that allow passage of currents with zero voltage, and when voltage is applied to them, they exhibit oscillations with a frequency to voltage constant of $\sim$484 MHz $\cdot$ μV. The resistive shunting causes the Josephson junctions to work in a nonhysteretic mode, which is necessary for low noise operation (9). An example of a thin-film dc SQUID, consisting of a square washer and Josephson junctions near the outside edge is shown in Fig. 1b (12,13). The usual symbol used to represent a dc SQUID is shown in Fig. 1c.

The SQUID ring (or washer) must be coupled to the external world and to the electronics that operates it (see Fig. 2a). When the dc SQUID is current biased, its *I–V* characteristics is similar to that of a nonhysteretic Josephson junction and the critical current I_0 is modulated

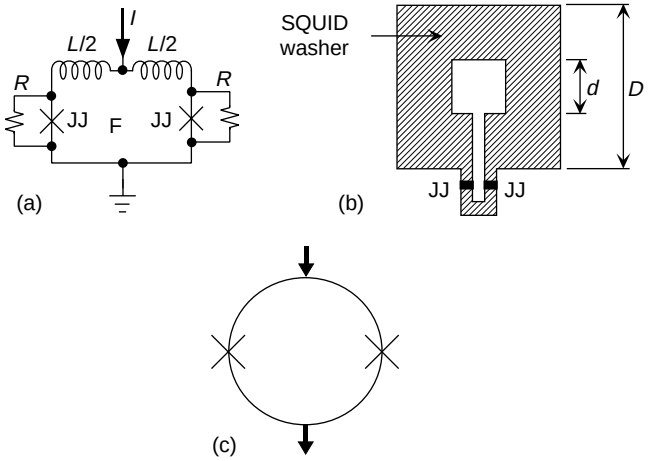


Figure 1. Thin-film dc SQUID. (a) Schematic diagram indicating inductances of the SQUID ring and shunting resistors to produce nonhysteretic Josephson junctions. (b) Diagram of a simple SQUID washer with Josephson junctions near the outer edge. (c) Symbolic representation of a dc SQUID, where the Josephson junctions are indicated by 'x'. (Reproduced with permission from Ref. 10).

by magnetic flux externally applied to the SQUID ring. The modulation amplitude is roughly equal to Φ_0/L (9), where Φ_0 is the flux quantum with magnitude $\sim 2.07 \times 10^{-15}$ Wb and L is inductance of the SQUID ring. The critical current is maximum for applied flux $\Phi = n\Phi_0$ and minimum for $\Phi = (n + 1/2)\Phi_0$. For monotonically increasing flux the average SQUID voltage oscillates as in Fig. 2d with period equal to 1 Φ_0. The SQUID transfer function is periodic (Fig. 2d) and to linearize it, the SQUID is operated in a feedback loop as a null detector of magnetic flux (14). Most SQUID applications use analogue feedback loop whereby a modulating flux with $\pm 1/4$ Φ_0 amplitude is applied to the SQUID sensor through the feedback circuitry (Fig. 2a,b). The modulation, feedback signal, and the flux transformer output are superposed in the SQUID, amplified, and demodulated in a lock-in detector fashion. The demodulated output is integrated, amplified, and fed back as a flux to the SQUID sensor to maintain its total input close to zero. The modulation flux superposed on the dc SQUID transfer function is shown in Fig. 2d and the modulation frequencies are typically several hundreds of kilohertz.

For satisfactory MEG operation, the SQUID system must exhibit large dynamic range, excellent interchannel matching, good linearity, and satisfactory slew rates. The analogue feedback loop is not always adequate and the dynamic range can be extended by implementing digital integrator as shown in Fig. 2c, and by utilizing the flux periodicity of the SQUID transfer function (15). The dynamic range extension works in the following manner: The loop is locked at a certain point on the SQUID transfer function and remains locked for the applied flux in the range of ± 1 Φ_0, Fig. 2d. When this range is exceeded, the loop lock is released and the locking point is shifted by 1 Φ_0 along the transfer function. The flux transitions along the transfer function are counted and are merged with the signal from the digital integrator to yield 32 bit dynamic range. This "flux slipping" concept can also be implemented using four-phase modulation (16), where the feedback loop jumps by $\Phi_0/2$ and can also provide compensation for the variation of SQUID inductance with the flux changes.

Flux Transformers

The purpose of flux transformers is to couple the SQUID sensors to the measured signals and to increase the overall magnetic field sensitivity. The flux transformers are superconducting and consist of one or more pickup coil(s) that are exposed to the measured fields. The pickup coil(s) are connected by twisted leads to a coupling coil that inductively couples the measured flux to the SQUID ring (as

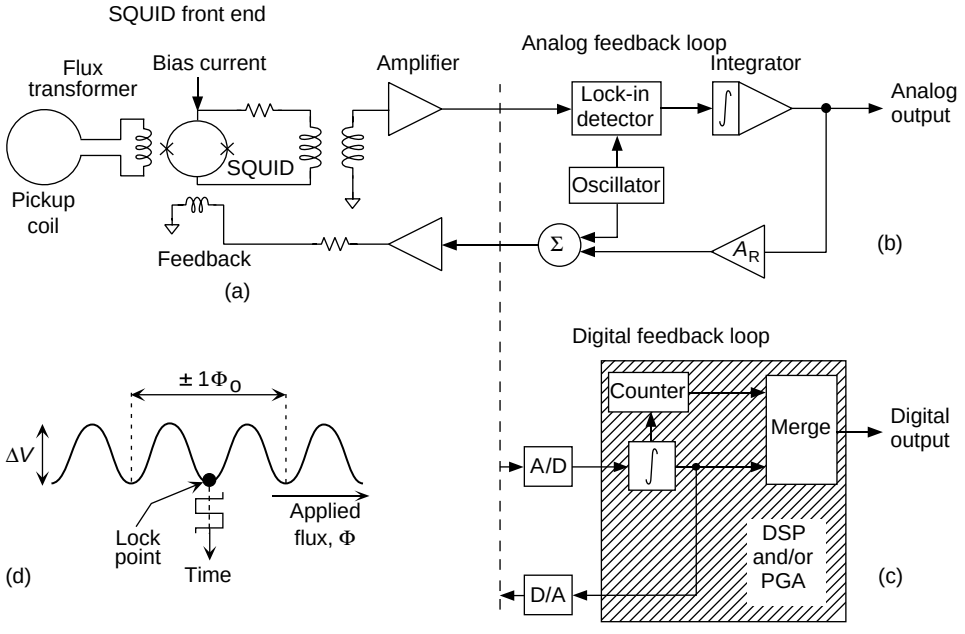


Figure 2. Examples of SQUID electronics, where the SQUID is operated as a null detector. (a) SQUID sensor is coupled to an amplifier. (b) Analogue feedback loop. (c) Digital feedback loop using digital signal processor (DSP) or a programmable logic array (PGA). (d) Feedback loop modulation. (Adapted with permission from Ref. 10).

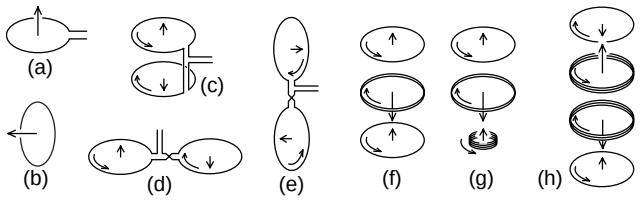


Figure 3. Examples of hardware flux transformers for biomagnetic applications. It is assumed that the scalp surface is at the bottom of the figure, (a) Radial magnetometer; (b) tangential magnetometer; (c) radial first-order gradiometer; (d) planar first-order gradiometer; (e) radial gradiometer for tangential fields; (f) second-order symmetric gradiometer; (g) second-order asymmetric gradiometer; (h) third-order gradiometer. (Reproduced with permission from Ref. 10).

shown in Fig. 2a). Because the flux transformers are superconducting, their gain is noiseless and their response is independent of frequency. The flux transformer pickup coil can have diverse configurations as shown in Fig. 3. A single loop of wire acts as a magnetometer and is sensitive to the magnetic field component perpendicular to its area, Fig. 3a and b. Two magnetometer loops can be combined with opposite orientation and connected by the same wire to the SQUID sensor. The loops are separated by a distance b and such a device is called a first-order gradiometer Fig. 3c–e, and the distance b is referred to as gradiometer baseline. The magnetic fields detected at the two coils are subtracted and the gradiometer acts as a spatial differential detector (this differential action is comparable to differential detection of electric signals (e.g., in electroencephalography, EEG). Fields induced by distant sources will be almost completely canceled by a gradiometer because both its coils will detect similar signals. On the other hand, near sources will produce markedly different fields at the two gradiometer coils and will be detected. Thus the gradiometers diminish the effect of the environmental noise that is typically generated by distant sources while remaining sensitive to near sources (e.g., neural sources). Similarly, first-order gradiometers can be combined with opposing polarity to form second-order gradiometers (Fig. 3f,g) and second-order gradiometers can be combined to form third-order gradiometers, (Fig. 3h). The flux transformers in Fig. 3 are called hardware flux transformers, because they are directly constructed in hardware by interconnecting various coils.

The main types of flux transformers used in commercial practice as the primary sensors are magnetometers (Fig. 3a), radial gradiometers (Fig. 3c), and planar gradiometers (Fig. 3d). These different sensor types will measure different spatial pattern of magnetic flux when placed over a current dipole as shown in Fig. 4. The radial magnetometer produces a field map with one maximum and one minimum, symmetrically located over the dipole with zero field measured directly above the dipole (Fig. 4a). The radial gradiometer in Fig. 4b produces similar field pattern as the magnetometer, except that the pattern is spatially tighter since it subtracts two field patterns measured at different distances from the dipole. The planar gradiometer field patterns are quite different from that of the

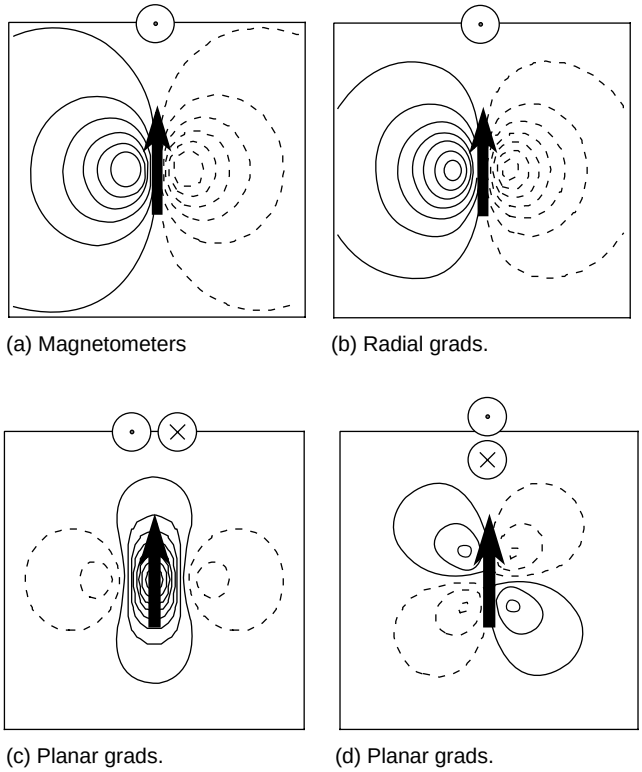


Figure 4. Response to a point dipole of several flux transformer types. A tangential dipole is positioned 2 cm deep in a semi infinite conducting space bounded by $x_3 = 0$ plane and its field is scanned by a flux transformer with its sensing coil positioned at $x_3 = 0$. Dipole position is indicated by a black arrow. Dimensions of each map are 14×14 cm. Schematic top view of the flux transformers is shown in the upper part of each figure. Solid and dashed lines indicate different field polarities. (a) Radial magnetometer; (b) radial gradiometer with 4 cm baseline; (c) planar gradiometer with 1.5 cm baseline aligned for maximum response; (d) planar gradiometer with 1.5 cm baseline aligned for minimum response. (Reproduced with permission from Ref. 10).

radial devices. If the two coils of the planar gradiometer are aligned perpendicular to the dipole, as in Fig. 4c, the planar gradiometer exhibits a peak directly above the dipole; if the two coils were aligned parallel to the dipole, the planar gradiometer exhibits a weak, clover-leaf pattern. When two orthogonal planar gradiometers are positioned at the same location, their two independent components can determine orientation of the current dipole located directly under the gradiometers (17).

In the absence of noise, there are no practical differences between these types of flux transformers. However, in the presence of noise, the signal-to-noise ratios (SNR) can differ greatly, resulting in significant performance differences between devices. For MEG applications, the magnitude of both the detected brain signal and environmental noise increases with increasing gradiometer baseline (distance between coils). Since the signal and noise functional dependencies on baseline are different, SNR exhibits a peak corresponding to an optimum baseline of ~3–8 cm for first-order radial gradiometers (10). Magnetometers can be thought of as gradiometers with very long baseline

and are not optimal because they can be overly sensitive to environmental noise. Planar gradiometers have good SNR for shallow brain sources but are suboptimal for deeper sources due to their short baselines resulting in poor depth sensitivity. Too long a baseline can also result in greater sensitivity to noise sources arising from the body itself, such as the magnetic field of the heart that may then contaminate the MEG signal. A detailed comparison of gradiometer design and performance can be found in (10).

Noise Cancelation

Introduction. Since biomagnetic measurements must be made in real world settings, the influence of noise on the measurements is a major concern in the design of biomagnetic instrumentation. Environmental noise affects biomagnetometer systems even when they are operated within shielded rooms. Environmental noise results from moving magnetic objects and currents (cars, trains, elevators, power lines, etc.). These noise sources are many orders of magnitude larger than signals of biomagnetic origin as shown in Fig. 5a. Note also, that only SQUID magnetometers have sufficient sensitivity for measuring biomagnetic signals of interest [atomic magnetometers are not yet suitable for biomagnetic applications (19)]. For MEG applications, the resolution or white noise level of the sensors should be much less than the "noise" level of brain activity ($\sim$30 fT $\cdot$ Hz$^{1/2}$). An example of background brain activity is shown in Fig. 5b. Also, certain MEG signal interpretation methods require the white noise to be as low as possible, however, the noise level cannot be made lower than the contribution of noise from the cryogenic vessel (dewar) itself. As a compromise, the majority of the existing MEG systems exhibit intrinsic noise levels of $<$10 fT $\cdot$ Hz$^{1/2}$ (typically $\sim$5 fT $\cdot$ Hz$^{1/2}$), yet are able to tolerate unwanted environmental noise many orders of magnitude greater.

Magnetic Shielding. Magnetic shielding is the most straightforward, though most costly method for reduction of environmental noise. A variety of shielded rooms have been used for biomagnetic applications and their relative shielding performance is shown in Fig. 6. The simplest shielding is accomplished through eddy currents by using a thick layer of high conductivity metal (20). Eddy current shielding is not effective at low frequencies, and therefore shielded rooms utilize high permeability μ-metal, which depending on the number of layers, can provide attenuation in the range from $\sim$30 to $\sim 10^5$ (21–24). Low frequency attenuation of nearly 10^8 was demonstrated with a whole-body, high Tc superconducting shield (25).

Environmental noise can also be reduced by active shielding, which can be employed either in unshielded environments (26), or in combination with shielded rooms (24,27,28). Active shielding system consists of a reference magnetometer, feedback electronics, and a set of compensating coils. The references measure the environmental noise and provide a signal that is amplified and fed into the

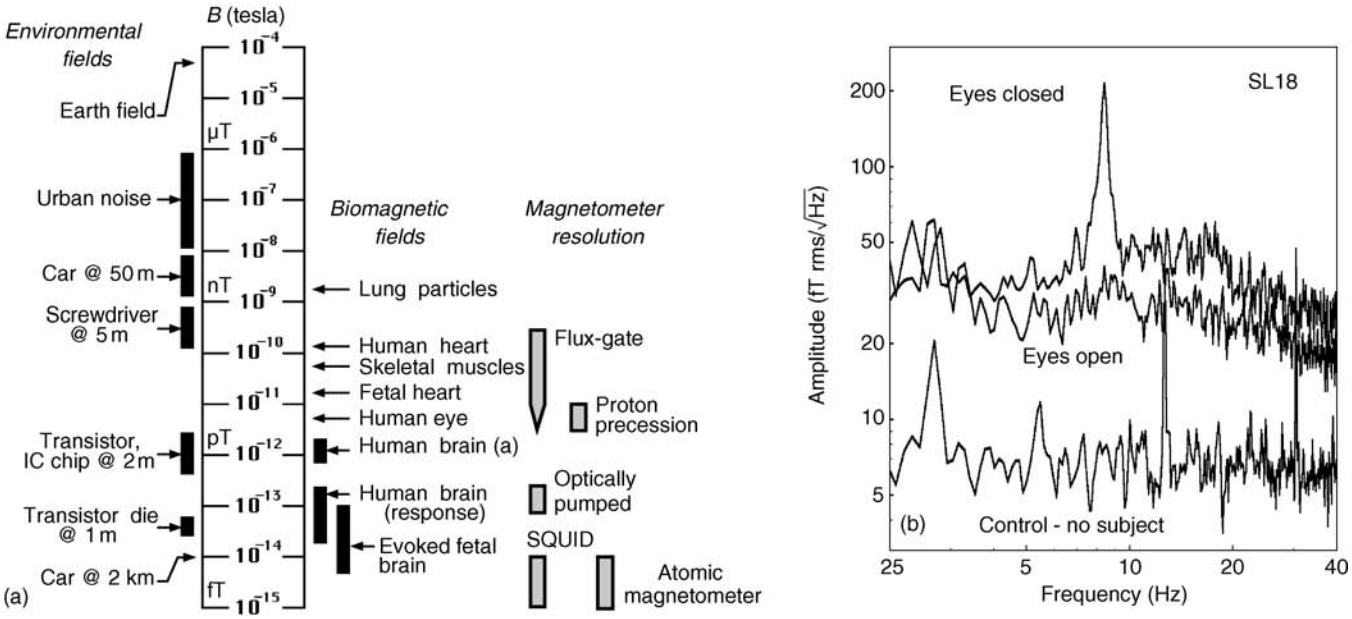


Figure 5. Environmental and brain generated noise. (a) Comparison of biomagnetic fields, environmental noise, and sensitivity in 1 Hz bandwidth of various types of magnetometers. (b) Spontaneous brain activity and the system noise measured in an unshielded environment, noise cancelation by synthetic third-order gradiometer, primary sensors are radial first-order gradiometers with 5 cm baseline. Control trace was collected with no subject in the helmet, large lines correspond to signals due to nearby rotating machinery. Eyes closed and open were collected with the subject in the MEG helmet. The presence of alpha activity (peak at 8 Hz) is visible in the eyes closed condition. (Reproduced with permission from Ref. 18).

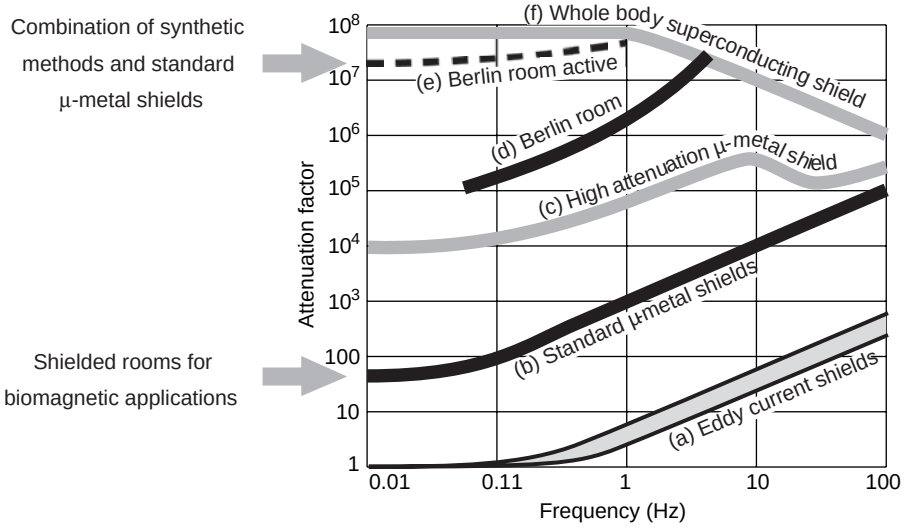


Figure 6. Noise attenuation of various shielded rooms as a function of frequency. (a) Eddy current Al rooms. (b) Standard μ-metal rooms used for MEG applications. (c,d) High attenuation μ-metal rooms. (e) Combination of high attenuation m-metal room in "d" and active shielding. (f) Whole-body high temperature superconducting shield. (Adapted with permission from Ref. 18).

compensating coils to reduce the noise. In general, the active shielding reduces the magnetic field noise due to far field sources and is effective only for magnetometers with no noise cancelation, while it has only a small effect on first-order gradiometers or magnetometers with noise cancelation. For higher order gradiometers, active shielding actually degrades system performance since the active coils can produce higher order gradients that are larger than that of the environmental noise.

Noise Reduction Using Higher Order Gradients. Since hardware noise cancelation (shielding or active noise cancelation) is usually not sufficient, additional methods, implemented in software or firmware, are employed. These methods either incorporate information from additional reference sensors or operate directly on the primary sensors. The reference sensors are typically a combination of SQUID magnetometers and gradiometers and the noise is canceled by synthesizing either higher order gradiometers or adaptively minimizing noise. The principle of synthetic gradiometer operation is similar for all gradiometer orders, and the method is illustrated for first-order gradiometer synthesis in Fig. 7a (29). The primary magnetometer detects the magnetic field component parallel to its coil normal, **p** (unit vector). The three reference magnetometers are orthogonal and their vector output, **r**, corresponds to the environmental field at the reference location, $\mathbf{r} \approx \mathbf{B}$. Then, if α_p is the primary magnetometer gain and α_r the reference gain (identical for all three references), the synthetic first-order gradiometer, $g^{(1)}$, can be derived as

$$g^{(1)} = m_p - \frac{\alpha_p}{\alpha_r}(\mathbf{p} \cdot \mathbf{r}) \approx \alpha_p \mathbf{p} \cdot \mathbf{G} \cdot \mathbf{b} \qquad (1)$$

where **b** is the gradiometer baseline (a vector connecting the primary sensor and the reference centers), and **G** is the first gradient tensor at the coordinate origin. Equation 1 states that the synthetic first-order gradiometer is a projection of the first-gradient tensor to the primary magnetometer orientation, **p**, and the baseline, **b**. To synthesize a second-order gradiometer, a primary hardware or synthetic first-order gradiometer, and a tensor first-gradient reference are used (Fig. 7b). Similar to Eq. 1, it can be shown that the synthetic second-order gradiometer output is a projection of the second gradient tensor to the coil orientation **p** and the first- and second-order gradiometer baselines $\mathbf{b}_1$ and $\mathbf{b}_2$. Synthesis of third- and higher order gradiometers is similar (29).

Adaptive methods can also be applied in addition to the synthetic gradiometers and can incorporate the same references as the gradiometers, but their coefficients are explicitly computed to minimize correlated noise (29). The advantage of synthesizing higher order gradiometers is that their coefficients are universal, independent of the noise character or sensor orientation (18). In contrast, the coefficients determined to adaptively minimize background noise are not universal because they depend on the noise character and sensor orientations (18) and assume that the noise environment is unchanging.

The noise cancelation achieved by various methods is illustrated in Fig. 8. The upper trace (a) shows the magnetic field noise outside a shielded room; and trace (b) shows the field noise after attenuation by the shielded room. The difference of the two slopes is due to the

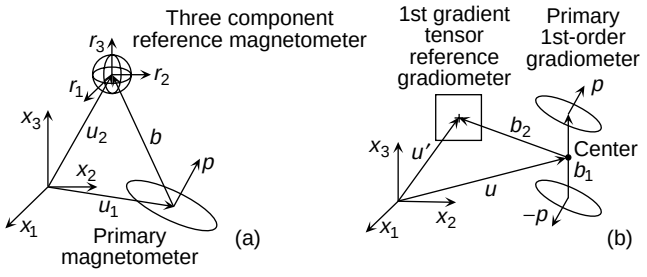


Figure 7. An illustration of gradiometer synthesis. (a) Synthesis of a first-order gradiometer from a primary magnetometer sensor and a vector magnetometer reference. (b) Synthesis of a second-order gradiometer from hardware first-order gradiometer and a first-gradient tensor reference. (Adapted from Ref. 30).

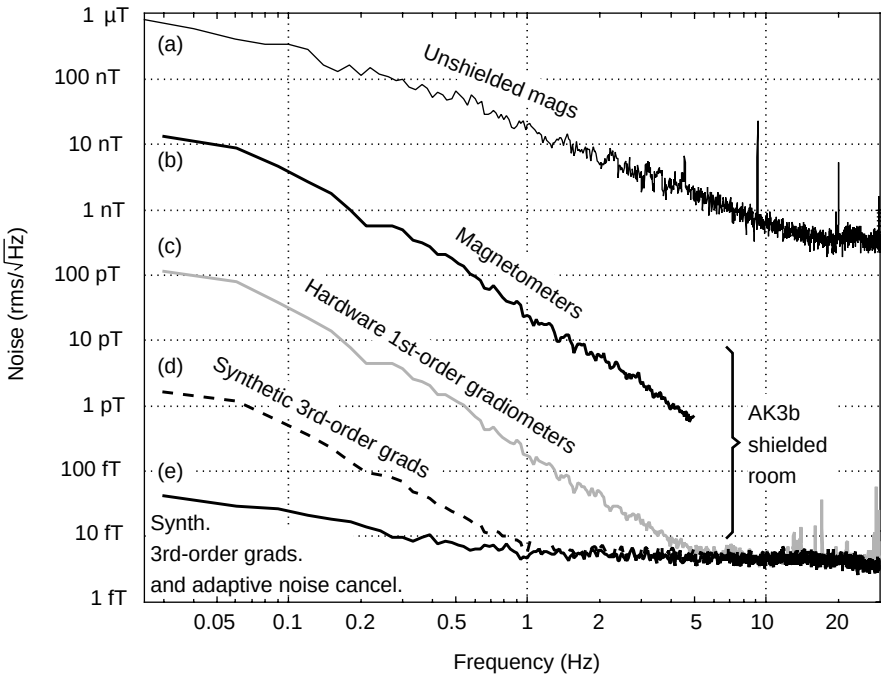


Figure 8. Reduction of environmental noise by a moderately shielded room, synthetic gradiometers, and adaptive methods. (a) Magnetic field noise outside a shielded room. (b) Field noise after attenuation by the shielded room. (c) Noise reduction by hardware first-order gradiometer with 5 cm baseline. (d) Noise reduction by synthetic third-order gradiometer (nearly four orders of magnitude lower noise than that of a shielded magnetometer in "b"). (e) Noise reduction by addition of adaptive methods to synthetic third-order gradiometer. (Adapted from Ref. 31).

frequency dependent eddy current shield that is part of the room. Hardware first-order radial gradiometers with 5 cm baseline reduce noise by nearly a factor of 100; and (c) a synthetic third-order gradiometer; (d) reduces the noise by almost another factor of 100. The low frequency environmental noise can further be reduced by adaptive method (e). The combination of all methods in Fig. 8 achieves attenuation of $>10^7$ at low frequencies.

Additional noise reduction methods can be employed in systems with a large number of channels. The simplest method is spatial filtering using Signal Space Projection (SSP) (32–34), which projects out from the measurement the noise components oriented along specific spatial vectors in signal space. The method works best when the signal and noise subspaces are nearly orthogonal. Related to SSP is noise elimination by rotation in signal space (35), which avoids loss of degrees of freedom encountered in SSP. These methods are discussed further in the Signal Interpretation section. More recently Signal Space Separation (SSS) has been proposed as a noise cancelation method in MEG (36). This approach was first proposed by Ioannides et al. (37) and reduces environmental noise by retaining only the "internal" component of the spherical expansion of the measured signal. This method can be applied to a number of problems inherent in biomagnetic measurements, including environmental noise reduction and motion compensation.

Cryogenics

The sensing elements of a biomagnetometer system (SQUIDs, flux transformers, and their interconnections) are superconducting and must be maintained at low temperatures. Since all commercial systems use low temperature superconductors, they must be operated at liquid He temperatures of 4.2 K. These temperatures can be achieved either with cryocoolers or by a cryogenic bath in contact with the superconducting components. The cryocoolers are attractive because they eliminate the need for periodic refilling of the cryogenic container. However, because they contribute magnetic and electric interference, vibrational noise, thermal fluctuations, and Johnson noise from metallic parts (38), they are not yet commonly used in MEG instrumentation. Present commercial biomagnetometer systems rely on cooling by liquid He bath in a nonmagnetic vessel with an outer vacuum space also referred to as a Dewar. An example of how the components may be organized within the Dewar for an MEG system is shown in Fig. 9a (39). The primary sensing flux transformers are positioned in the Dewar helmet area. The reference system for the noise cancelation is positioned close to the primary sensors and the SQUIDs with their shields are located at some distance from the references, all immersed in liquid He or cold He gas. The Dewar is a complex dynamic device that incorporates various forms of thermal insulation, heat conduction, and radiation shielding, as shown Fig. 9b. Most commercial MEG and MCG systems have reservoirs holding up to 100 L of liquid He and can be operated for periods of several days before refilling. An excellent review of the issues associated with the Dewar construction is presented in (38).

Biomagnetometer Systems: Overview

Even though magnetic fields have been detected from many organs, so far the most important application of biomagnetism has been the detection of neuromagnetic activity of the human brain. This interest led to the development of sophisticated commercial MEG systems. The current generation of these systems consists of helmet shaped multisensor arrays capable of measuring activity

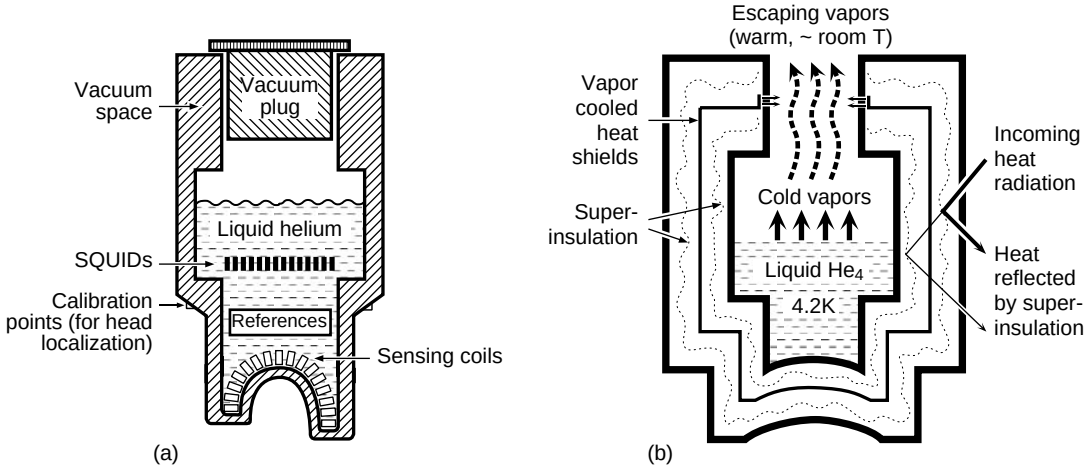


Figure 9. Schematic diagram of cryogenic containers used for whole-cortex MEG. (a) Placement of various MEG components relative to the cryogenic Dewar. (b) Principles of the Dewar operation. Reproduced with permission from (10).

simultaneously from the entire cerebrum. In contrast, multichannel magnetocardiogram (MCG) systems consist of a flat array of radial or vector devices (40–45) or systems with a smaller number of channels operating at liquid N_2 temperatures (46–51) for better placement over the chest directly above the heart. These flat array systems can also be placed over other areas of the body to measure peripheral nerve, gastrointestinal, or muscle activity. These systems can even be placed over the maternal abdomen to measure heart and brain activity of the fetus and a custom shaped multichannel array specifically designed for fetal measurements has recently been introduced (39,52).

MEG Systems. A diagram of a generic MEG system is shown in Fig. 10. The SQUID sensors and their associated flux transformers are mounted within a liquid He dewar suspended in a movable gantry to allow for supine or seated patient position. The patient rests on an adjustable chair or a bed. All signals are preamplified and transmitted from the shielded room to a central workstation for real-time acquisition and monitoring of the magnetic signals. At present, the majority of MEG installations use magnetically shielded rooms, however, progress is being made toward unshielded operation (18,40). The MEG measurements are often complemented by simultaneous EEG measurements or peripheral measures of muscle activity or eye movement. Most MEG installations have provisions for stimulus delivery in order to study brain responses to sensory stimulation and video and intercom systems in order to interact with the patient from outside the shielded room. Multichannel MEG systems are commercially available from a number of manufacturers (39,53–56).

For MEG localization of brain activity to be useful, particularly in clinical applications, it must be accurately known relative to brain anatomy. The anatomical information is usually obtained by magnetic resonance imaging

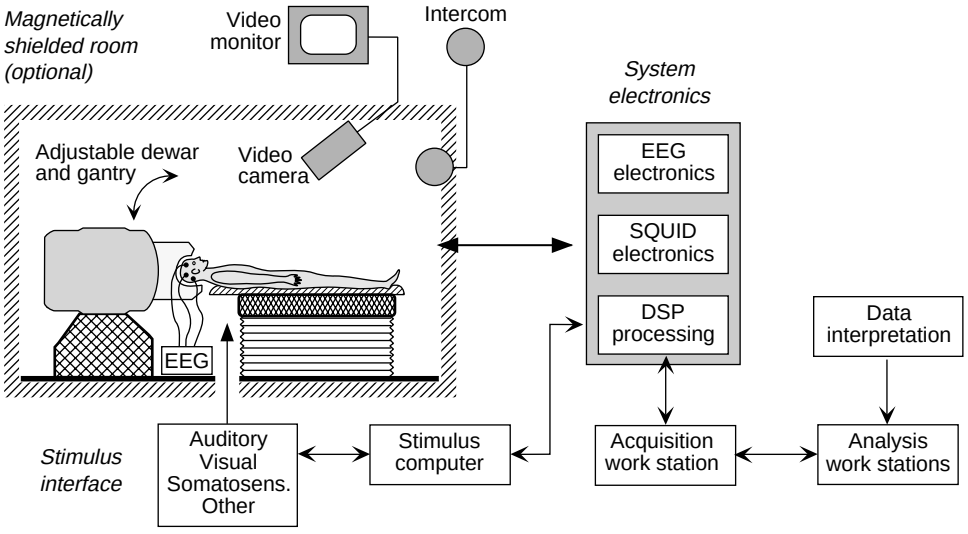


Figure 10. Schematic diagram of a typical MEG installation in a magnetically shielded room. (Reproduced with permission from Ref. 10).

(MRI), and the MRI images are required during the MEG interpretation phase. The registration of the MEG sensors to the brain anatomy is performed in two steps. First, the head position relative to the MEG sensor array is determined in order to accurately position MEG sources within a head-based coordinate system. Second, the head position relative to the MRI anatomical image is determined to allow transfer of MEG sources to the anatomical images. There are different methods for such registration. The simplest one uses a small number of anatomical markers positioned on identical locations on the head surface that can be measured both by MEG and MRI (e.g., small coils for MEG and lipid contrast markers for MRI) usually placed at anatomical landmarks near the noise and ears (18). To improve localization accuracy, the head shape can be digitized in the MEG coordinate system by a device mounted on the dewar (57) or by the MEG sensors (10). The surface of the head can also be constructed from segmented MRI and the transformation between the two systems can be determined by alignment of the two surfaces (58–60).

Biosusceptometers. A somewhat different system design is encountered in biomagnetometer systems used for the measurement of magnetic materials in the human body, such as iron content in the liver or magnetic contaminants in the lung. These instruments contain both SQUID sensing coils and a superconducting magnet operated in persistent mode. The system is suspended over the patient's body on a bed with a waterbag placed between the patient and dewar to provide continuity of the diamagnetic properties of body tissue. Figure 11 illustrates the layout of a biosusceptometer system for liver measurements with a patient in a supine position on a moveable bed. The patient is moved vertically relative to the SQUID gradiometer-magnet system and flux changes due to the susceptibility of the liver are monitored. These measures of magnetic moment can then be used to estimate the concentration of the paramagnetic compounds within the liver (62–64).

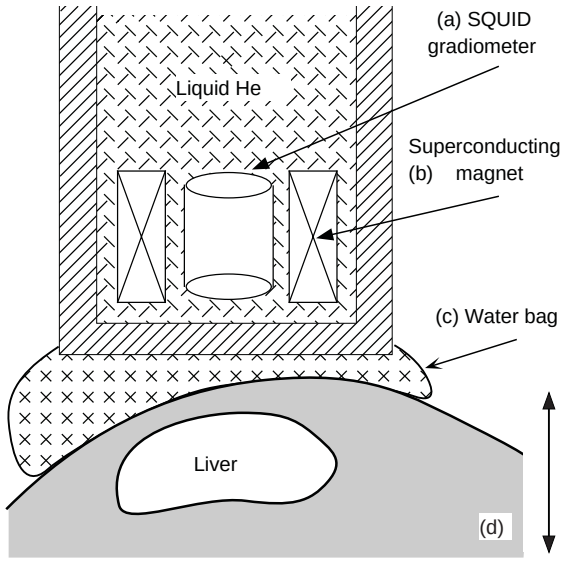


Figure 11. Schematic diagram of a liver susceptometer. (a) SQUID gradiometer. (b) Superconducting magnet. (c) Bag filled with water to simulate the diamagnetism of human body tissue. (d) Patient on a bed that is vertically movable. (Reproduced with permission from Ref. 61).

Signal Interpretation

Biomagnetometers measure the distribution of magnetic field outside of the body. Although the observed field patterns provide some information about the underlying physiological activity, ideally one would like to invert the magnetic field and provide a detailed image of the current distribution within the body. Such inversion problems are nonunique and ill defined. The nonuniqueness is either physical (65) or mathematical due to being highly underdetermined (i.e., there are many more sources than sensors). In order to determine the current distribution, it is necessary to provide additional information, constraints, or simplified mathematical models of the sources. The field of source modeling in both MEG and MCG has been an intensive area of study over the last 20 years. In the following section we shall review briefly various methods of source analysis as it is applied to MEG, although these methods apply to other biomagnetic measurements such as MCG, with the main difference being the physical geometry of the conductor volumes containing the sources. For detailed reviews of mathematical approaches used in biomagnetism (see 66–69).

Neural Origin of Neuromagnetic Fields. Magnetic fields of the brain measured by MEG are thought to be the primarily due to activation of neurons in the gray matter of the neocortex, whereas action potentials in the underlying fiber tracts (white matter) have been shown to produce only poorly synchronized quadrupolar sources associated with weak fields (70,71). Some subcortical structures have also been shown to produce weak yet measurable magnetic fields, but are difficult to detect without extensive signal processing (72,73). The generation of magnetic fields in the human brain is illustrated in Fig. 12. The neocortex of the brain (shown in Fig. 12a) contains a large number of pyramidal cells arranged in parallel (Fig. 12b) that in their resting state maintain an intracellular potential of ca. -70 mV. Excitatory (or inhibitory) synaptic input near the cell body or at the superficial apical dendrites results in the flow of charged ions across the cell membrane producing a graded depolarization (or hyperpolarization) of the cell. This change in polarization results in current flow inside the cell, called *impressed* current and corresponding return or *volume* currents that flow through the extracellular space in the opposite direction. Studies carried out in the early 1960s (74,75) demonstrated that these extracellular or volume currents are main generators of electrical activity measured in the electroencephalogram or EEG. The combination of excitatory and inhibitory synaptic inputs to different cortical layers can produce a variety of sink and source patterns through the depth of the cortex, each associated with current flow along the axes of elongated pyramidal cells toward or away from the cortical surface.

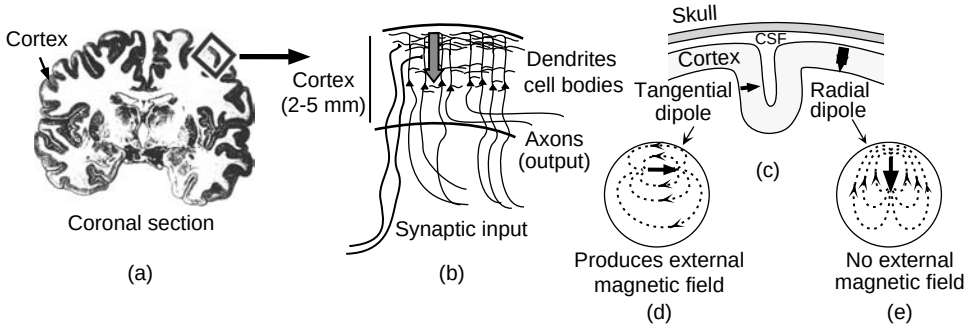


Figure 12. Origin of the MEG signal. (a) Coronal section of the human brain. The neocortex is indicated by dark outer surface. (b) Pyramidal cells in the cortex have vertically oriented receptive areas (dendrites). Depolarization of the dendrites at the cortical surface due to excitatory synaptic input results in Na^+ ions entering the cell producing a local current source and a current sink at the cell body, resulting in intracellular current flowing toward the cell body (arrow). (c) The cortex has numerous sulci and gyri resulting in currents flowing either tangentially or radially relative to the head surface. (d) Tangential currents will produce magnetic fields that are observable outside the head if modeled as a sphere. (e) Radial currents will not produce magnetic fields outside of the head if modeled as a sphere. (Adapted from Ref. 10).

Synchronous activity in large populations of these cells summate to produce the positive and negative time-varying voltages measured at the scalp surface in the EEG (76).

Okada et al. (77) carried out extensive studies over the last 20 years on the neural origin of evoked magnetic fields using small array "microSQUID" systems to measure directly magnetic fields from *in vitro* preparations in the turtle cerebellum and mammalian hippocampus. These studies have shown that although both extracellular and intracellular currents may contribute to externally measured magnetic fields, it is primarily intracellular or impressed currents flowing along the longitudinal axis of pyramidal cells that are the generators of evoked magnetic fields. A recent review of this work is presented in (78). Note that, since MEG measures mainly intracellular currents and EEG the return volume currents, the pattern of electrical potential over the scalp due to an underlying current source will reflect current flow in opposite direction to that of the magnetic field, as has been demonstrated in physical models (79) and human brain activity (80). In addition, activation of various regions of the enfolded cortical surface (the gyri and sulci) will result in current flow that is either radial or tangential to the scalp surface, respectively (Fig. 12c). If the brain is modeled as a spherical conducting volume, then due to axial symmetry it can be shown that only the tangential currents will produce fields outside the sphere (81) (Fig. 12d and e). Using *in vivo* preparations in the porcine brain, it has been experimentally demonstrated that, in contrast to the EEG, magnetic fields are relatively undistorted by the presence of the skull, and are generated primarily in tangentially oriented tissue (78). It has been recently shown, however, that MEG is insensitive only to a relatively small percentage of the total cortical surface in humans due to this tangential constraint (82). There is some uncertainty as to the extent of cortical activation typically measured by MEG. Current densities in the cortex have been estimated to be on the order of 50 $\text{pA} \cdot \text{m} \cdot \text{mm}^2$(83) suggesting that cortical areas of at least 20 mm^2 must be activated in order to produce a sufficiently large external field to be observed outside the head (66,68). However, current densities as high as 1000 $\text{pA} \cdot \text{m} \cdot \text{mm}^2$ have been recorded *in vitro* (77) indicating that much smaller areas of activation may be observed magnetically.

Equivalent Current Dipoles. The equivalent current dipole or ECD (81,84) is the oldest and most frequently used model for brain source activity. It is based on the assumption that activation of a specific cortical region involves populations of functionally interconnected neurons (macrocolumns) within a relatively small area. When measured from a distance, this local population activity can be modeled by a vector sum or "equivalent" current dipole that represents the aggregate activity of these neurons. The ECD analysis proceeds by estimating a priori the number of equivalent dipoles and their approximate locations, and then adjusting the dipole parameters (location and orientation) by a nonlinear search that minimizes differences between the field computed from the dipole model and the measured field (Fig. 13). This can be done at one time sample, or it can be extended to a time segment, where several dipoles are assumed to have fixed positions in space, but variable amplitude. Such models are referred to as "spatiotemporal" dipole models (85). The dipole fit procedures require the calculation of the magnetic field produced by a current dipole at each sensor: also termed the *forward solution*. Since the frequency range of interest for biomagnetic fields is <1 kHz, the quasistatic approximations of Maxwell's equations apply. If the head is assumed to be approximately spherical in shape it can be represented by a uniformly conducting sphere and the radial magnetic field of an ECD with magnitude $\mathbf{q}$, is given by the radial component of the well-known Biot–Savart law, $B_{\text{rad}}(\mathbf{r}) = \mathbf{B}(\mathbf{r}) \cdot \mathbf{r}/|\mathbf{r}|$, where the Biot–Savart

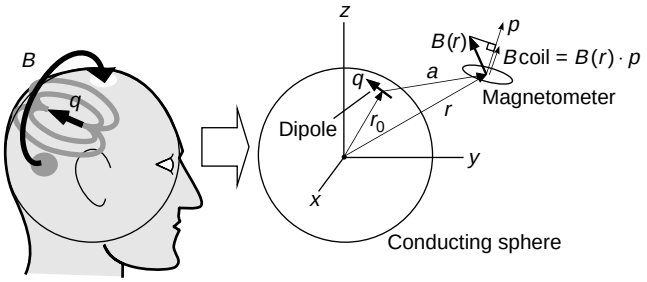


Figure 13. Magnetic fields due to an equivalent current dipole source will exit and reenter the head that can be modeled as a spherical shaped conducting medium. Calculation of the field magnitude ($\mathbf{B}_{coil}$) measured by a magnetometer coil due to a current dipole $\mathbf{q}$ at location $\mathbf{r_0}$ inside a sphere is given by the projection of the calculated vector field $\boldsymbol{B(r)}$ onto the direction normal to the surface area of the coil indicated by the unit vector $\mathbf{p}$, such that $\boldsymbol{B}_{coil} = \boldsymbol{B}(\boldsymbol{r}) \cdot \boldsymbol{p}$. The orientation of $\mathbf{q}$ is assumed to be tangential to the sphere surface. For gradiometer devices, the measured output of the gradiometer can be calculated as the difference between the field magnitudes calculated separately at each of the coils.

vector field, $\mathbf{B}(\mathbf{r})$, is given by

$$\mathbf{B}(\mathbf{r}) = \frac{\mu_0}{4\pi} \frac{\mathbf{q} \times (\mathbf{r} - \mathbf{r}_o)}{|\mathbf{r} - \mathbf{r}_o|^3} \tag{2}$$

where $\mathbf{r}_o$ is the ECD position and $\mathbf{r}$ is the position where the field is measured. For multiple ECDs or continuously distributed sources, Eq. 2 will also include the sum over all sources or the integral over the volume of the conducting sphere.

Generally, the vector of the external magnetic field is produced by both the primary current density reflecting the impressed (intracellular) currents, and volume currents that produce "secondary sources" on the surface of the volume conductor. For complex shapes, the calculation of the external field also requires knowledge of the conductivity profile of the conducting volume. The assumption of spherical symmetry, however, simplifies the calculation, and the vector field $\mathbf{B(r)}$ due to a current dipole $\mathbf{q}$ in a sphere at location $\mathbf{r}_o$ (Fig. 13b) is given by Sarvas (81) as

$$\mathbf{B}(\mathbf{r}) = \frac{\mu_0}{4\pi F^2} \{F\mathbf{q} \times \mathbf{r}_o - [(\mathbf{q} \times \mathbf{r}_o) \cdot \mathbf{r}]\nabla F\} \tag{3a}$$

where

$$F = a(ra + r^2 - \mathbf{r}_o \cdot \mathbf{r}) \tag{3b}$$

and

$$\nabla F = (r^{-1}a^2 + a^{-1}\mathbf{a} \cdot \mathbf{r} + 2a + 2r)\mathbf{r} - (a + 2r + a^{-1}\mathbf{a} \cdot \mathbf{r})\mathbf{r}_o \tag{3c}$$

and $\mathbf{a} = \mathbf{r} - \mathbf{r}_o$, $a = |\mathbf{a}|$, $r = |\mathbf{r}|$ and the permeability of free space $\mu_0 = 4\pi \times 10^{-7}$ H · m. The sensing coil measures the component of the vector field $\mathbf{B(r)}$ perpendicular to its surface area as shown in Fig. 13b. If the field is measured only in the radial direction, Eq. 3 simplifies to the radial component of the Biot–Savart law (Eq. 2), and the volume currents do not contribute any field. It can be seen from Fig. 13 that the definition of the origin of the theoretical sphere relative to the head will influence the calculation of the external magnetic field and thus plays a significant role in the accuracy of the single sphere approach. Since the head is not perfectly spherical, improved accuracy of the forward solution can be achieved by using more realistic models of the conducting surfaces and boundary element methods for the calculation of the magnetic field (69), but these methods are more computationally demanding. A simple improvement over the single sphere model can be achieved by using a multiple-sphere model, where independent spheres are determined for each sensor by evaluating local head curvature in the sensor vicinity (86).

The ECD procedure is very sensitive to the SNR and dc offsets, and therefore works best when applied to averaged brain responses that are well time locked to a sensory or motor event and requires an accurate estimate of signal baseline (e.g., prestimulus activity). This approach has proven useful for modeling simple patterns of focal brain activity, yet is compromised by interaction or "cross-talk" between simultaneously active sources, requiring that the number of dipoles be correctly specified. Also, ECD models do not correctly describe spatially "extended" sources: areas of cortical activity that may extend over an area of several square centimeters.

Minimum Norm. The dipole model assumes that the brain activity is localized in one or several small areas of the brain. Sometimes it is required to obtain a more general solution without an a priori assumption about the source distribution. This can be obtained by minimum norm methods, first proposed for MEG by Hämäläinen and Ilmoniemi (84). This inverse problem is underdetermined, solutions are diffuse, and the unweighted minimum norm favors solutions close to the sensors. The minimum norm method has subsequently been adapted to produce more localized solutions. The algorithm, FOCal Undetermined System Solution (FOCUSS) utilizes a recursive linear estimation based on weighted pseudoinverse solution (87) and the Minimum Current Estimate (MCE) utilizes the L1-norm approach (88). A related method, Magnetic Field Tomography (MFT) (89) utilizes weights and a regularization parameter that are optimized according to the given experimental geometry and noise. Another minimum norm-based method is the algorithm LORETA (LOw Resolution Electromagnetic TomogrAphy) (90). This algorithm introduces a spatial second derivative operator (Laplacian) into the weighting function and seeks the minimum norm solution subject to the maximum smoothness condition. This requirement is justified on a physiological assumption that neighboring points in the brain are likely to be synchronized. The method produces low spatial resolution that is a consequence of the smoothness constraint. Methods based on simulated (surrogate) data have also been proposed for producing distributed, unbiased solutions based on the minimum norm (91).

Bayesian Inference. Bayesian inference has also been applied to the biomagnetic inverse problem, using probability distributions of many possible source solutions. This approach can easily incorporate a priori information that may influence the likelihood of features of the current

distribution based on anatomy, maximum current strength, smoothness, and so on (92,93). This method determines expectation and variance of the a posteriori source current probability distribution given source prior probability distribution and data set (94,95). The model can include probability weightings determined from other imaging techniques such as functional MRI (fMRI) or positron emission tomography (PET) to influence the MEG current images.

Signal Space Projection. Signal space projection. (33,34) and beamformers are spatial filters that can separate signal from noise on the basis of their relationship in signal space (a M-dimensional space, where M is the number of MEG channels). The application of spatial filtering to MEG was first proposed by Robinson and Rose (96). This original article sparked growing interest in spatial filtering by the MEG community that still continues. The spatial filtering depends on the assumption that component vectors corresponding to different neuronal sources have distinct and stable (fixed) directions in signal space, and only their magnitudes are functions of time. If the vectors are defined by modeling the field produced by known dipole sources, SSP can be used as a spatial filter that passes only signals corresponding to these known sources. Thus, we can define the output of a spatial filter as $y_\theta(t) = \mathbf{P}_\parallel \mathbf{m}(t)$, where $\mathbf{P}_\parallel$ is the parallel projection operator (95) constructed from the forward solutions of the dipole source(s) of interest, and $\mathbf{m}(t)$ represents a vector of instantaneous MEG measurement at time t. The output of the spatial filter then provides a time series that is the estimate of changing strength of the dipole source(s) over time. Alternatively, if the vectors associated with artifact patterns are known, SSP can be used to remove these artifacts from the signal using orthogonal projection operators (32). If the signal vectors are determined from patterns in the data, the source model need not even be known. Note that restricting all sources to current dipoles in a known volume conductor model reduces SSP to a multiple dipole approximation (34).

Beamformers. The SSP method does not separate well sources that are not in orthogonal subspaces. To overcome this limitation, source analysis can be done by beamforming (borrowed from radio-communication and radar work). Beamformers utilize spatial and temporal correlations to obtain information about uncorrelated dipolar sources. The Linearly Constrained Minimum Variance (LCMV) beamformer in the form now used in MEG analysis was first described in 1972 (97) and can be used without specific information about source orientation. An introduction to the beamformers may be found in (98) and a relatively recent review of various beamforming techniques in (99). As in the case of SSP, if vector $\mathbf{m}(t)$ represent an instantaneous MEG measurement in M-dimensional space, we can define a spatial filter centered on the location "θ" as $y_\theta(t) = \mathbf{W}_\theta^T m(t)$, where $\mathbf{W}_\theta$ is a weight matrix. Only tangential sources contribute to the MEG signal. They can be decomposed into two orthogonal tangential directions and the corresponding forward solutions, $\mathbf{B}_{\theta 1}$ and $\mathbf{B}_{\theta 2}$, can be arranged in a forward solution matrix as $\mathbf{H}_\theta = [\mathbf{B}_{\theta 1}, \mathbf{B}_{\theta 2}]$. The beamformer weights are determined by minimizing the power projected from the location , $P_\theta = \mathbf{W}_\theta^T \mathbf{C} \mathbf{W}_\theta$, subject to the unity gain condition, $\mathbf{W}_\theta^T \mathbf{H}_\theta = \mathbf{I}$, where $\mathbf{C}$ is the covariance matrix of the measurement and $\mathbf{I}$ is the identity matrix. The weights are given as (100)

$$\mathbf{W}_\theta = \mathbf{C}^{-1}\mathbf{H}_\theta(\mathbf{H}_\theta^T \mathbf{C}^{-1}\mathbf{H}_\theta)^{-1} \tag{4}$$

An alternative approach known as synthetic aperture magnetometry (SAM) defines an optimal dipole orientation for each spatial filter location (101). Only one vector is retained, $\mathbf{H}_\theta = \mathbf{B}_\theta$ simplifying Eq. 6 to $\mathbf{W}_\theta = \mathbf{C}^{-1}\mathbf{B}_\theta (\mathbf{B}_\theta^T \mathbf{C}^{-1} \mathbf{B}_\theta)^{-1}$. This approach produces higher spatial resolution due to less projected sensor noise by the spatial filter (102). The beamformer weights can be used to compute the time course of the dipole magnitude variation or power at a single location in the brain independently of other active sources, provided sources are not highly correlated. An especially useful quantity is the normalized power $Z_\theta^2 = P_\theta / N_\theta$, where $N_\theta^2 = \mathbf{W}_\theta^T \Sigma \mathbf{W}_\theta$ is the sensor noise projected by the beamformer from location 'θ', and Σ is the sensor noise covariance matrix (100). In contrast to P_θ and N_θ, the parameter Z_θ^2 behaves gracefully through the center of the model sphere and does not exhibit a singularity. A spatial image of brain activity can be obtained by computing the normalized power at individual brain voxels, θ, one at a time over a region of interest.

Multiple Signal Classification. MUltiple SIgnal Classification (MUSIC) is a signal space scanning method and is related to beamforming (103,104). MUSIC requires an initial nonlinear step of partitioning the data covariance matrix into signal and noise subspaces using standard eigendecomposition methods. This partitioning can be more readily determined from the averaged data and as a result the method is more difficult to apply to spontaneous brain activity. Sources are located by scanning of the brain volume and at each location requiring that the dipole forward solution be orthogonal to the noise subspace (or parallel to the signal subspace). A more recent implementation known as recursively applied and projected MUSIC (RAP-MUSIC) projects out each located source and then repeats the scanning procedure (105). Similar to beamforming, MUSIC also assumes there are fewer sources than sensors, the sources are uncorrelated and the noise is white. In the limit of high SNR (e.g., averaged data), a small number of sources, and white noise, the MUSIC localizer function and beamformer based source power estimates differ only by a scaling factor.

Principal Component Analysis. Principal Component Analysis (PCA), for example (106,107), also determines the signal and noise subspaces. The method is based on second order statistics and attempts to fit dipoles into the orthogonal principal spatial vectors of the singular value decomposition of the data. For mixtures of components corresponding to nonorthogonal spatial vectors, the PCA cannot account for the structure of the data (108). The PCA has been shown to be potentially inaccurate, as it can mislocalize dipoles even in noiseless simulations.

Independent Component Analysis. Independent Component Analysis (ICA) is a relatively new technique that

allows separation of sources that are linearly mixed at the sensors. The method is also called blind source separation, because the source signals are not directly observed and nothing is known about their mixture (109,110). The method uses higher order statistics and in realistic situations is often more successful than PCA (108). The mixing model used for the separation is usually stated as $\mathbf{m}(t) = \mathbf{A}\mathbf{s}(t)$, where $\mathbf{m}(t)$ is the instantaneous vector of the measurement, $\mathbf{s}(t)$ is the instantaneous source activity vector, and $\mathbf{A}$ is the mixing matrix. The procedure provides solution for an unmixing matrix $\mathbf{B}$, such that the estimated source activity is given as $\hat{\mathbf{s}}(t) = \mathbf{B}m(t)$, where $\hat{\mathbf{s}}$ is the estimate of the source vector $\mathbf{s}$. The sources are assumed to be statistically independent and the separation is obtained by optimizing a contrast function, that is, a scalar measure of some distributional property of the output $\hat{\mathbf{s}}$. The contrast functions are based on entropy, mutual independence, high order decorrelations, and so on. The ICA has been applied to MEG and EEG to either remove artifacts or extract desired signals (111,112).

APPLICATIONS OF BIOMAGNETIC MEASUREMENTS

Magnetoencephalography: Basic Studies

The most prevalent and rapidly growing application of biomagnetism is the field of magnetoencephalography (MEG): the measurement of human brain activity. This field of basic and clinical research is also referred to as *neuromagnetism* or *magnetic source imaging* (113–117). The latter term is often used to refer to the localization of neural sources with respect to individual brain anatomy by the combination of MEG source modeling with structural imaging techniques such as MRI (see Magnetic Resonance Imaging). As noted in the introduction, the first magnetic fields recorded from the human brain involved the observation of spontaneous alpha rhythm activity. This was soon followed by the application of MEG measurements to the study of *evoked responses* of the human brain: Time-averaged responses to discrete sensory or motor events that provide sufficient SNR to allow for the localization of brain regions contributing to the externally measured field patterns. Due to its excellent time resolution and ability to measure neuronal function directly, MEG has continued to generate interest within the field of human neuroscience as a complement to other methods of functional brain imaging based on metabolic or hemodynamic changes in brain function, such as fMRI (see Magnetic Resonance Imaging) or PET (see Positron Emission Tomography). Present MEG practice includes measurement of both evoked and spontaneous signals and is used for clinical purposes and for investigation of a wide range of brain processes.

Somatosensory Evoked Fields. Evoked responses to stimulation of the human somatosensory system (somatosensory evoked fields or SEFs) were first reported by Brenner and colleagues in 1987 (118). The observed magnetic brain response to electrical stimulation of the digit was of great interest since it demonstrated a well-characterized dipolar field pattern over the scalp indicative of a single neural generator located in the underlying somatosensory cortex. Subsequent studies have shown that early components of the SEF occurring at latencies of 20–50 ms reflect early activation of the primary somatosensory cortex contralateral to the side of stimulation, and are generally well modeled as single ECD source in these brain regions [see (119) for a recent review]. The earliest component at a poststimulus latency of 20 ms (sometimes referred to as the "M20" or "N20m" since it is considered the magnetic equivalent of the negative N20 potential measured in the EEG) arises from the posterior bank of the central sulcus: a primary somatosensory projection area. By stimulating different body parts, it can be shown that the N20m source reflects the somatotopic or "homuncular" organization of the ascending neural pathways of the somatosensory system to this brain region (120). Figure 14 shows a typical SEF response to stimulation of the median nerve at the wrist and the localization of an ECD model fit to the N20m source in the corresponding somatosensory cortex.

When using extensive signal averaging, MEG recordings also reveal low amplitude high frequency oscillations in the 300–900 Hz range during the period of the N20m

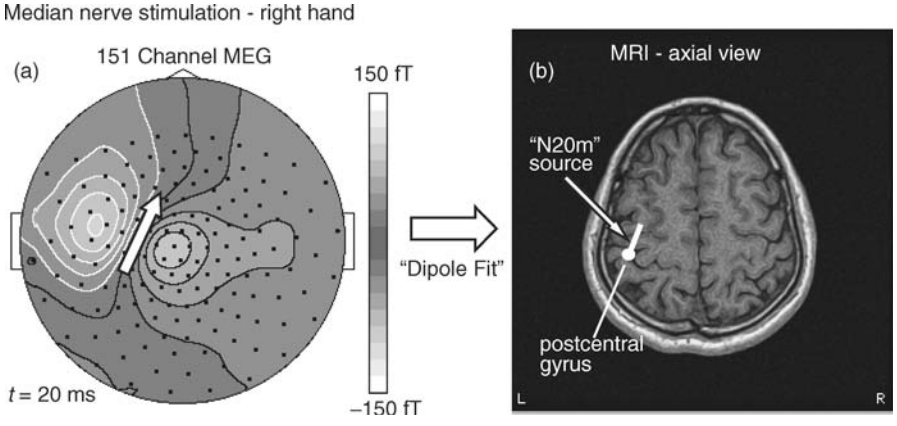


Figure 14. (a) Topographic map (polar projection with nose upwards) of the magnetic field pattern recorded from 151 MEG channels over the scalp at a latency of 20 ms following stimulation of the right median nerve (average of 600 stimuli). White contours indicate outgoing fields and solid contours, ingoing fields. Arrow indicates direction of current flow below the scalp corresponding to the dipolar field pattern over the left hemisphere. (b) Location of a single ECD source corresponding to the magnetic field pattern shown in (a) indicated by white dot with tail indicating direction of current flow superimposed on an axial slice of the individual's MRI. Location is in the hand region of the primary somatosensory cortex.

response (121). These oscillations have been proposed to reflect the activity of inhibitory interneurons in the somatosensory cortex (122) although cortico-thalamic pathways have also been shown to play a possible role (123). The N20m is followed by reversals of the same pattern at latencies of 30 and 40 ms that appear to reflect additional activation of somatosensory areas. These are followed by more complex and widespread activity from ~80 to 150 ms after stimulation that reflects bilateral activation of secondary somatosensory areas in the parietal operculum and is most likely related to higher order processing of somatosensory input (124). The MEG responses at latencies of 50–70 ms are elicited by mechanical stimulation of the digits (125) and reflect somatotopically organized sources in the primary somatosensory cortex (80). A number of MEG studies have used mechanical SEFs to demonstrate functional reorganization or "plasticity" of the somatosensory cortex resulting from anesthetic block or damage to the peripheral nerves or amputation (126), or even as the result of musical training (127).

Movement Related Fields. The first recordings of magnetic fields accompanying simple finger movements were reported in the early 1980s. Deecke et al. (128) observed slow magnetic field changes over sensorimotor areas of the brain preceding voluntary movements of the digits. These "readiness fields" begin approximately a half a second prior to the onset of a voluntary movement and are thought to represent activation of brain areas involved in motor preparation (129). Dipole source analysis suggests that premovement fields arise primarily from bilateral activation of the primary motor cortex (even for unilateral movements) with larger amplitude fields and dipole magnitudes the contralateral to the side of movement (130).

Movement-evoked fields (MEFs) accompany the onset and execution phase of simple movements. The first component (MEFI) is the largest in amplitude and begins ~100 ms after onset of EMG activity in the involved muscles. These responses appear to arise from sources in the postcentral gyrus, most likely reflecting sensory feedback to cortex from proprioceptors in the muscles (131) and are correlated with movement velocity (132). Movements made in response to a sensory cue show a very similar pattern of activity, but with a shorter latency of onset of premovement activity (133). Passive movements also elicit magnetic responses thought to reflect activation of the proprioceptive inputs to areas of the postcentral gyrus (134). MEG mapping studies have demonstrated activity in motor cortex during motor imagery providing evidence of the involvement of these brain areas in the simple imagination of movement (135).

MEG–EMG Coherence. By using a single channel magnetometer, Conway et al. (136) made the interesting observation of increased coherence (correlation in the frequency domain) between the surface electromyogram (EMG) in a contracting muscle and MEG recordings made over the contralateral motor areas. Subsequently, there has been a great deal of interest in the relationship between MEG–EMG coherence and the functional relationship between spontaneous cortical rhythms and EMG activity during movement (137). Interestingly, changes in the frequency of coherence varies with the strength of muscular contraction and recent studies have shown that MEG–EMG coherence may reflect the underlying physiology of tremor in patients with Parkinson's disease (138) or essential tremor (139).

Sensorimotor Rhythms. The MEG studies have also provided evidence for the functional significance of specific oscillatory brain activity in humans associated with both somatosensory stimulation and motor output. These centrally distributed rhythms were first observed in the EEG, and are predominant at frequencies ~10 Hz (also referred to as the mu rhythm) and in the range of 20–30 Hz. The MEG studies have been able to show that these are functionally independent cortical oscillations that originate from postcentral and precentral regions, respectively (140,141). These sensorimotor rhythms are suppressed during median nerve stimulation, followed by a transient increase or "rebound" of 20–30 Hz rhythms within 500 ms after stimulus onset. A similar pattern of suppression followed by rebound is observed during voluntary movements (142). These rhythmic changes are modulated by sensorimotor tasks such as movement or passive tactile stimulation and motor imagery or even observation of another individual's movements (140). Rhythmic activity is not amenable to the same signal averaging technique used for evoked fields and therefore the ECD source modeling approach is more difficult to apply. Spatial filtering methods, however, provide a new approach to the localization of frequency dependent power changes in cortical areas using MEG and have been applied to the localization of rhythmic changes induced by somatosensory stimulation (141,143) and voluntary movements of the digits (144).

Visual Evoked Fields. One of the first magnetic evoked responses recorded from the human brain was the visual evoked field or VEF reported by Brenner et al. in 1975 (145). Robust responses can be elicited at latencies of 100–150 ms following visual stimulation using light flashes or visual pattern contrast changes (e.g., reversing checkerboard stimuli). However, early VEF responses pose a challenge in terms of modeling their sources due to the complex enfolded cross-like shape of the primary (striate) visual cortex: also referred to as the "cruciform" model. More recently, investigators have successfully modeled early VEF components in primary visual cortex by stimulating restricted portions of the total visual field using both monochrome (146) and color (147) pattern stimuli and have produced source configurations that reflect the retinotopic organization of the primary visual cortex. Due to the difficulty in applying ECD models to the VEF response, spatial filtering methods such as beamforming have been found to be useful for imaging visual cortex function (148). Figure 15 shows the activation of primary visual cortex by a steady state visual pattern (reversing checkerboard) using the SAM beamforming algorithm. Visual stimuli also activate several nonprimary (extrastriate) visual areas depending on the attributes of the stimulus. A number of MEG studies have shown activation of brain areas related to higher order visual processes such as detection

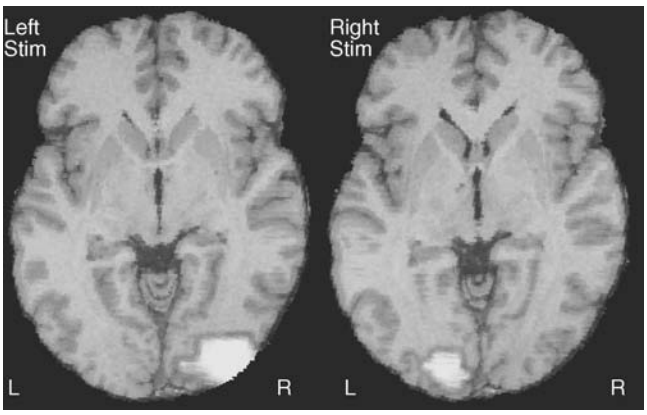


Figure 15. Images of following response to flickering checkerboard stimulus at $f = 17$ Hz presented to the left or right visual field, using a whole-head MEG recordings and the synthetic aperture magnetometry (SAM) beamformer algorithm. The image shows increased source power as yellow (lighter) colored areas at the posterior portion of the brain, corresponding to increased power at 34 Hz ($2f$) in the primary visual cortex of the contralateral hemisphere. Unpublished data. (Courtesy of K. Singh.)

of coherent motion (149,150). Clinical applications of VEFs have been limited, although abnormal VEFs have been reported in cases of strabismic amblyopia (151).

Auditory Evoked Fields. Auditory evoked fields were first reported by Reite et al. (152) and subsequent MEG mapping studies have demonstrated that responses at latencies of 50 and 100 ms reflect activity of primary auditory cortex in the temporal lobes (153). The largest response occurring ~100 ms following stimulus onset, termed the M100, has been the most extensively studied auditory evoked field response. The M100 is bilateral for both binaural and monaural stimuli and has been shown to reflect the frequency specific (tonotopic) organization of the primary auditory cortex (154). These magnetic evoked responses are of interest in the study of the functional organization of the auditory system as they reflect perceptual attributes of auditory stimulation such as perceived pitch or the frequency profiles of complex speech sounds (155,156). Auditory responses to repetitive (steady-state) auditory stimuli show enhanced amplitude in the EEG at presentation rates of ~40 Hz and were initially thought to represent volume conducted thalamic responses. However, MEG steady-state auditory responses were shown by Weinberg et al. (157) to reflect oscillations at the stimulus frequency in the auditory cortex. Subsequent studies have suggested that 40 Hz auditory responses reflect thalamocortical networks in the brain responsible for integration of sensory input (158). The 40 Hz MEG response has recently been used to measure temporal integration times in the primary auditory cortex (159).

MEG Studies of Higher Brain Function. Although early MEG studies have provided useful information regarding the early processing of sensory input and motor output, one of the more intriguing potential uses of MEG is the noninvasive study of higher brain function. The EEG studies of cognitive function have been carried out using event-related potentials (ERPs) for many decades. The MEG measurements using similar paradigms have helped gain a better understanding of the neural basis of many ERP components. Early MEG studies have had some success in measuring brain responses related to short-term memory (160), target detection tasks (161), or selective attention (162). More recently, the use of whole head MEG systems have enabled the study of more complex aspects of cognitive processing in humans, such as face recognition (163) and object naming (164). Basic research on brain mechanisms related to speech and language is also a promising area of application for MEG. Early studies have shown that speech processing is affected by incongruous visual feedback at the level of the auditory cortex (165). More recent studies have attempted to localize magnetic brain responses related to syntactic (166) and semantic (167) language processes as well as the processing of speech sounds (168). The MEG responses have also been used to study abnormal processing of sensory input during reading in dyslexic children (169) and during speech in stutterers (170).

A great deal of progress has been made in studying higher brain function with MEG by applying traditional source analysis methods to ERP components. However, these complex brain processes often involve activation of multiple brain regions complicating the interpretation of the data in terms of simple ECD models. Moreover, many of these processes may not be highly time-locked to specific sensory or motor events. More recent approaches have focused on oscillatory brain activity and synchronization or "phase-locking" between different cortical areas. Accordingly, this has produced increased interest in brain imaging methods with fine temporal resolution such as MEG. Rhythmic activity in the so-called gamma frequency band (30–90 Hz) is of particular interest since it is associated with cognitive processes such as feature binding within a sensory modality that may underlie perception (171). Recent studies have also described changes in neuromagnetic rhythms associated with observation and imitation of other individual's actions that appear to originate in brain areas associated with learning through imitation (172). Since changes in spontaneous brain rhythms are not necessarily time-locked to a stimulus onset, alternative signal processing techniques are required (173). The combination of spatial filtering source analysis methods and time frequency and phase analysis may be particularly well suited to measure these aspects of higher order brain function (141) and constitute a new and interesting avenue of research in human cognition.

Magnetoencephalography: Clinical Applications

Presurgical Functional Mapping. One of the more prevalent clinical applications of MEG is the localization of so-called "eloquent cortex" (those areas that subserve sensory, motor, speech, and memory function) prior to neurosurgery in order to prevent loss of these functions as a result of the surgical procedure. Due to displacement cortical tissue by space occupying lesions such as tumors, or natural variability in cortical morphology, identification of these brain

areas may not be possible by visual inspection alone and can be aided by functional localization of these areas using MEG. This is achieved by activating primary sensory areas associated with visual, somatosensory and auditory stimulation, and applying ECD models to the early evoked response—a method generally referred to as *presurgical functional mapping* (114,116). For example, the N20m source of the somatosensory evoked field can be consistently and reliably localized in most individuals and used as an estimate of the location of the central sulcus prior to surgical removal of brain tissue in the region of the primary motor or somatosensory cortex (174).

Determination of the language dominant hemisphere is also necessary prior to surgical resection of cortical tissue near language areas of the temporal lobe. This is routinely done through highly invasive procedures such as selective anesthesia of the left and right hemispheres (*Wada test*) or direct cortical stimulation intraoperatively. The use of MEG for the localization of brain areas that are specific to the processing of speech, as distinct from areas associated with the simple processing of auditory input, constitutes a challenging area of research, however, some recent progress has been made in this area (175–177). In addition to using MEG source imaging to identify functional and pathological brain areas in surgical planning, these functional data can also be incorporated into frameless stereotaxic *neuronavigation* systems. These systems allow surgeons to identify the corresponding brain areas in the functional and structural images during the surgical procedure. The MEG-based neuronavigation is rapidly becoming a useful clinical tool for the surgical treatment of epilepsy, tumors and other brain disorders (114,178,179).

Epilepsy. Due to its high temporal resolution and ability to localize focal brain activity, there has been a long interest in the application of MEG to the study of epilepsy. In many case, intractable seizures can be controlled by the removal of the epileptogenic zone: brain tissue from which seizure activity originates. The identification of this area may be aided by the measurement of abnormal electrical activity between seizures. These interictal (between seizure) spikes arise from an *irritative zone* that may be correlated with the epileptogenic zone in cases of focal epilepsy (180). The localization of ECD sources based on MEG recordings of interictal spiking activity has been shown to be highly correlated with the localization of this zone as identified by other methods such as direct intracranial monitoring from depth or subdural electrode grids [for recent reviews see (181–184)]. Interictal spikes are generally of much larger amplitude than sensory evoked responses and ECD models can be used to localize individual spike events without averaging. However, the area activated may be quite large and exhibit a high degree of spatial variability, and as a result the aggregate locations or clusters of many spike sources are often used to estimate the putative location of the irritative zone. Other methods, such as spatial filtering by beamformers (SAM), are currently being investigated and may help overcome some of the limitations of the single ECD approach to the localization of epileptogenic areas. Even in cases where the precise location of the epileptogenic zone is not clearly identified, MEG may help to guide the placement of subdural grids, and in some cases may be used to evaluate the propagation of abnormal electrical activity between multiple brain regions. The diagnostic yield of MEG measurements of interictal activity varies with different forms of epilepsy and appears to be highest for neocortical epilepsy (185) and can also aid in the differentiation of different types of epilepsy (186).

Since the site of brain pathology may not be known in advance, particularly in nonlesional epilepsy, the introduction of whole-head MEG systems has drastically improved the feasibility of using MEG as a routine clinical procedure for presurgical epilepsy evaluation allowing the data to be acquired in a more rapid and standardized manner. The main drawbacks to the application of MEG in epilepsy is the inability to measure brain activity during or just prior to seizure onset due to head movement, and the difficulty in performing long-term monitoring of interictal activity, although this is somewhat ameliorated by the introduction of MEG systems that allow recording from patients in the supine position and while asleep. Although there is some debate on the overall usefulness of MEG in the presurgical evaluation of epilepsy (183) comparisons with other modalities such as EEG, functional MRI, and intracranial electrical recordings (114,182) indicate that MEG provides useful complementary, and in some cases unique information for the surgical treatment of epilepsy.

Other Clinical Applications. Although presurgical functional mapping and epilepsy have been the main areas of clinical application of MEG, other brain disorders have been studied. This includes the use of MEG to study changes in electrical brain activity associated with tumors that often manifests as abnormal low frequency activity (187) or other disturbances in brain rhythmic activity (188) and may help identify the functional integrity of surrounding brain tissue (189). Abnormal low frequency activity has been associated with other brain lesions such as those due to stroke and in epilepsy (190). The MEG has also been used to study recovery after stroke due to functional reorganization of the cortex (191) and its relationship to rehabilitation and outcome (192) and in the evaluation of patients with mild head injury (193). Low frequency neuromagnetic activity has been hypothesized to be an index of spreading cortical depression associated with migraine (194).

The MEG studies have focused on pain related brain responses by selectively stimulating the $A\delta$ and C fiber systems painful CO_2 laser stimulation of the skin (195) or direct electrical stimulation of nerve fibers (196). This type of somatosensory stimulation produces long latency responses in secondary somatosensory areas located in the parietal operculum, and insula: brain regions known to be involved in the perception of pain. Such studies are promising for the clinical treatment of chronic pain, although are challenging due to the difficulty in discriminating activation of brain areas due to painful versus nonpainful somatosensory input, and the invasiveness of the procedure.

More recently, MEG measures have also been combined with neurochemical imaging methods such as magnetic resonance spectroscopy (MRS). In these studies, correlations have been found between MEG activity and changes in levels of neurotransmitter and other brain metabolites in ischemia or in brain areas harboring tumors (197). Although still a new area of study, there is a great deal of interest in the application of MEG to psychiatric disorders. For example, MEG studies have reported abnormal auditory evoked magnetic fields in schizophrenic patients (198) and patients with Alzheimer's disease (199) or in individuals with developmental disorders such as autism (200).

Magnetocardiography

The first biomagnetic measurements in humans were measurements of the magnetic field of the heart. The field of magnetocardiography or MCG has not expanded as rapidly as that of MEG, although a number of research centers have continued to develop the MCG method for the noninvasive evaluation of cardiac disease. As described in the Instrumentation section, MCG requires instrumentation designed for the adequate sampling of the heart's magnetic field over the chest and a number of instruments have been developed and installed at research centers around the world, including systems based on high temperature SQUIDs. Source modeling based on magnetic field measurements is somewhat simplified in the case of MEG due the ability to model the head as a spherically shaped conductor, whereas, modeling of the electrical activity of the heart requires realistic models of the conducting properties of the thorax and its influence on the distribution of magnetic fields arising from the heart. As a result, source localization methods in MCG often employ boundary element methods for forward calculations (201). Source localization in MCG is further complicated by the continuous movement of the heart itself. Nevertheless, MCG has been successfully used in the diagnosis of cardiac disease. For recent reviews see (202–204).

Since the 1980s a number of studies have focused on the use of MCG for the 3D localization of the origins of abnormal electrical activity of the heart. This includes abnormal activity underlying cardiac arrhythmias, such as Wolff–Parkinson–White syndrome, which involves abnormal electrical activity (preexcitation) in the accessory pathway. Recent studies have shown that MCG studies provide more accurate localization of the site of pathology than standard multichannel electrocardiogram techniques (205). The identification of the generators of heart arrhythmias is useful in presurgical evaluation for interventional procedures such as catheter ablation therapy, or in the screening of patients at risk for ventricular tachycardia (202) or coronary artery disease (206). Another application of MCG is in the assessment of ischemic areas of the heart after infarction by the detection of regions of low current density (207).

Fetal Studies

One of the more intriguing new applications of biomagnetism is the noninvasive measurement of activity of the fetal heart and brain. Since the first report of the detection of an evoked response from the fetal brain in 1985 (208) there has been a great of interest in developing instrumentation for the measurement of biomagnetic fields from the human fetus. The main challenges for the measurement of fetal MCG or MEG is the detection of biomagnetic sources that are distant from the detector array, and the difficulty in establishing the position of the fetal heart and brain during measurement. The latter has been partially resolved by the combination of fetal biomagnetic measurements with 3D ultasound imaging and new instrumentation has been recently designed for optimum placement and sensitivity of the sensory array to detect fetal heart and brain responses.

Fetal MCG. The largest biomagnetic signal arising from the fetus is the fetal magnetocardiogram or fMCG. The first recording of fMCG was demonstrated in 1984 (209). The fMCG signal magnitude is quite large, but due to proximity of the fetus to the mother's heart, signal processing methods are required to first remove the large maternal heart signal, after which the P, QRS, and T segments of the fMCG can be discerned with high reliability in fetuses beyond the twentieth week of gestation (210). Fetal heart rate variability has been shown to be a good indicator of fetal well being (211). This method has been applied to the detection of fetal arrhythmias (212) and may provide a useful diagnostic or screening tool for fetal congenital heart defects (213), or for the assessment of fetal health in high risk pregnancies. An overview of fMCG can be found in (214).

Fetal MEG. Due to the distance of the fetal brain from the surface of the maternal abdomen, the fetal MEG (fMEG) signal is difficult to detect without high sensitivity biomagnetometer with large coverage, large number of channels, and optimal placement of the sensor array. In addition, the fetal brain signals are small in comparison with an adult and their measurement is performed in the presence of strong interference from the maternal and fetal heart signals and various abdominal signals (intestinal electrical activity, uterine contractions, etc.). Measurements of fetal brain responses to sensory stimulation are also hampered by the difficulty in delivering the sensory stimulus to the fetus. However, fetal auditory evoked responses have been successfully recorded by presenting high amplitude auditory stimuli directly to the mother's abdomen (215,216). In order to successfully eliminate the interference due to cardiac signals, which can be >100 times larger than the fMEG, the latter efforts employed various signal extraction methods (spatial filtering, PCA, etc.) in addition to averaging. Magnitudes of fMEG responses to transient tone bursts are in the range of $\sim$8–180 fT and the latencies range from $\sim$125 to nearly 300 ms, decreasing with the increasing gestation age (217). The response is typically observed in not more than about 50% of examined subjects. Fetal responses to steady-state auditory clicks have also been reported (218) as well as spontaneous fetal brain activity in the form of burst suppression (219). The strength of these signals can be

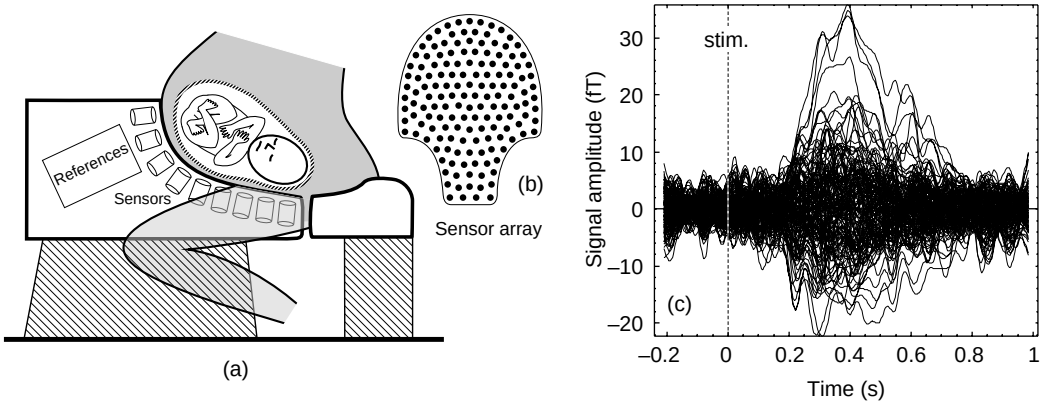


Figure 16. Dedicated system for fetal MEG measurement. (a) Schematic diagram of SQUID Array for Reproductive Assessment (SARA) (52). (b) Layout of 151 sensing channels. (c) Example of flash evoked fMEG response, overlay of 151 SARA channels. The fetus with gestation age of 28 weeks was stimulated by 33 ms duration flashes of 625 nm wavelength light (220). Vertical dashed line corresponds to the flash stimulus onset. (Adapted from Ref. 221).

relatively large, up to 500 fT, and can represent interference during evoked fMEG responses.

Early fMEG experiments used single or multiple-channel probes with relatively small area of coverage, requiring a search for the region with the largest signals. Recently, a dedicated fetal MEG system, the SQUID Array for Reproductive Assessment (SARA), was constructed and operated (52). The SARA system consists of an array of 151 SQUID sensors covering the mother's anterior abdominal surface in late gestation, from the perineum to the top of the uterus as shown in Fig. 16a and b. The primary sensor flux transformers are axial first-order gradiometers, with 8 cm baseline with a nominal SQUID sensor noise density of 4 $fT/Hz^{1/2}$. The SARA system is now being used routinely and was recently used to measure the first fetal visual evoked field to high intensity light stimuli presented to the maternal abdomen (220). An example of a flash evoked response from the fetal brain is shown in Fig. 16c.

Other Applications

Biosusceptometry. The greatest interest in biosusceptometry has stemmed from its potential to assess noninvasively iron overload in the human liver. This potentially fatal condition arises in individuals with hemoglobinopathies that require frequent blood transfusions (e.g., sickle cell anemia) or involve abnormal production of hemoglobin (hemachromatosis and beta-thalassemia). Standard methods for assessing iron overload can be highly invasive (e.g., liver biopsy) and biosusceptometry offers a safer and potentially more accurate diagnostic tool. Iron, which is normally strongly ferromagnetic, is stored in the liver bound by the proteins ferritin and hemosiderin and exhibits a strong paramagnetic response. As a result, measurement of the magnetic moment produced by placing the liver in a uniform magnetic field will be proportional to the total amount of iron in the liver: a method known as *biomagnetic liver susceptometry* (BLS). The basis for this technique was first proposed and the first measurements carried out in the late 1970s (222).

Most approaches to the measurement of hepatic iron concentration involve placing the patient's abdomen directly under a magnetic sensor that also contains a field coil that produces a magnetic field, and lowering the patient by a fixed distance to measure the change in field amplitude due to the magnetized liver (Fig. 11). In order to eliminate the effect of the surrounding air, a water-filled bellows is placed between the abdomen and the device to simulate the diamagnetic properties of the other tissues in the body. The main challenge to accurate estimates of hepatic iron content using BLS is the remaining effect of the varying susceptibility of the lungs and air filled compartments in the abdomen. Since this technique requires the application of a dc magnetic field to the body on the order of about 0.1 T, it is a much more invasive technology in comparison to MEG and MCG, and may be contraindicated in patients with implanted medical devices such as pacemakers. A detailed review of the clinical applications of BLS can be found in (223).

Peripheral Nerve Studies. It is known from the pioneering studies of Wikswo and colleagues (70) that the propagation of action potentials in nerve fibers produces quadrupolar like sources that have a rapidly diminishing magnetic field with distance. This is due to the fact that action potentials consist of a traveling wave of depolarization in the axon, followed closely by a wave of repolarization. In addition, due to varying conduction velocities in the peripheral nerves, action potentials in different axons will not necessarily summate to produce coherent synchronous activity. As a result, activation of compound nerve bundles does not produce coherent dipole-like sources as in the case of the neocortex. However, with sufficient signal averaging it is possible to record the magnetic signature of the sensory nerve action potentials noninvasively in the human: a technique

referred to as *magnetoneurography*. These measures have been achieved by placing single channel magnetometers or flat arrays of magnetic sensors over the peripheral nerve pathways and electrically stimulating the nerve. The predicted quadrapolar pattern of traveling actions potentials resulting from electrical stimulation of the finger was reported by Hoshiyama et al. (224) using a 12 channel "micro-SQUID" device placed over the wrist. Mackert et al. (225), using a 49 channel flat triangular array of first-order radial gradiometers were able to measure compound action potentials elicited by tibial nerve stimulation in sensory nerves entering the spinal cord at the lower lumbar region, and have recently using this method clinically to demonstrate impaired nerve conduction in the patients with S1 root compression.

Magnetopneumography. Magnetopneumography refers to the measurement of the remanent magnetism of ferromagnetic particles in the lungs. This technique may be used to assess lung contamination encountered in occupations that may involve the inhalation of ferromagnetic dust particles such as arc-welders, coalminers, asbestos, and foundry and steel workers. Similar to liver biosusceptometry, magnetopneumography involves the application of a weak dc magnetic field to the thorax. However, the field is applied for only a short interval in order to produce a remanent magnetization of ferromagnetic material, usually iron oxides such as magnetite. This remanent magnetic field is then measured to assess to total load of ferromagnetic particles in the lung. These measures can be used to evaluate the quantity and clearance rates of these substances (226,227). A related measure is *relaxation*: the decay of the remanent field due to the reorientation of the magnetic particles away from their aligned state after application of the dc field. Relaxation times are thought to reflect cellular processes in the lung associated with clearance or macrophage activity on the foreign particles. Recent studies have used magnetopneumography to study the effect of smoking on clearance times of inhaled magnetic particles (228).

Gastrointestinal System. Biomagnetic measurements have also been applied to other areas of the human body. The human gastrointestinal system produces electrical activity associated with the processes of peristalsis and digestion of food. For example, slow electrical activity at frequencies of $\sim$3 cycles/min (0.05 Hz) can be recorded from the human stomach using cutaneous surface electrodes or magnetically: a technique referred to as *magnetogastrography* (MGG). This activity arises from the smooth muscle of the stomach and the detection of changes in frequency with time has been proposed as a method of characterizing gastric disorders (229). Another novel application of biomagnetic instrumentation to gastrointestinal function, is the 3D tracking of the transport of magnetic materials through the gut. This technique has been termed *magnetic marker monitoring* (MMM) and can be used to monitor the passage and disintegration (by measuring decrease in magnetic moment) of magnetically labeled pharmaceutical substances through the gastrointestinal system (230).

FUTURE DIRECTIONS

Since its inception 40 years ago with the first recording of the magnetic field of the heart, the field of biomagnetism has expanded immensely to become a major field of basic and applied research. The field of magnetoencephalography, or MEG, has in recent years become a recognized neuroimaging technique, with the development of advanced instruments for the measurement of the electrical activity of the brain with exquisite temporal and spatial resolution. Biomagnetic instrumentation is now at a mature state, with commercially developed measurement systems available for a variety of biomagnetic applications. For example, whole head MEG systems are installed worldwide in $>$100 research laboratories and clinical centres and are now being used in routine clinical diagnostic procedures. Nevertheless, there remain many areas for further improvement of both instrumentation and data analysis approaches and techniques. In terms of instrumentation, biomagnetometer systems with increased number of sensing channels and capable of unshielded operation will likely be developed, and present systems that require frequent refilling with liquid Helium may be replaced by systems with longer hold times and less frequent cryogen replenishment. The latter may be accomplished either by incorporation of cryocoolers, or the use of sensors that do not require liquid He. The last two technical innovations, combined with production of larger numbers of MEG systems will also help reduce the cost of these instruments.

The analysis and interpretation of biomagnetic measurements is possibly the most significant area for continued research and development, and much progress has been made in the implementation of new signal processing algorithms for the extraction of biomagnetic signals, or improving the spatial resolution of source localization methods. There has been recent interest in combining MEG with its high temporal resolution and other functional imaging techniques such as functional MRI. In addition, advanced image processing techniques, such as the automated extraction of the cortical surface of the brain from structural MRI, will allow the use of more precise physical models of biomagnetic sources. Combination of MEG with its counterpart EEG may also help to develop more accurate models of brain activity. These advancements will aid the development of new clinical applications of biomagnetism such as the use of MEG to study psychiatric disorders, or to study the effects of drug treatments on brain processes related to cognitive deficits, or gain insight into the physiological mechanisms underlying various brain disorders in children, for example, learning disabilities, dyslexia, and autism. Finally, novel applications of biomagnetic measurements, for example, the measurement of heart and brain activity in the fetus, will lead to new applications of biomagnetism in clinical medicine and will further drive the development of improved technology. In sum, biomagnetism will continue

to grow as a novel and powerful noninvasive technique for the study of physiological processes in humans in both health and disease.

BIBLIOGRAPHY

Cited References

1. Baule GM, McFee R. Detection of the magnetic field of the heart. Am Heart J 1963;66:95–96.
2. Cohen D. Magnetoencephalography: evidence of magnetic fields produced by alpha-rhythm currents. Science 1968; 161:784–786.
3. Josephson BD. Possible new effect in superconductive tunneling. Phys Lett 1962;1:251–253.
4. Zimmerman JE. Recent developments in superconducting devices. J Appl Phys 1971;42:30–37.
5. Cohen D, Edelsack EA, Zimmerman JE. Magnetocardiograms taken inside a shielded room with a superconducting point-contact magnetometer. Appl Phys Lett 1970;16:278–280.
6. Cohen D. Magnetoencephalography: detection of the brain's electrical activity with a superconducting magnetometer. Science 1972;175:664–666.
7. Cheyne D, Vrba J, Crisp D, Betts K, Burbank M, Cheung T, Fife A, Haid G, Kubik P, Lee S, McCubbin J, McKay J, McKenzie D, Spear P, Taylor B, Tillotson M, Weinberg H, Basar E. Use of an unshielded, 64 channel, whole cortex MEG system in the study of normal and pathological brain function. Proceedings of the Satellite Symposium on Neuroscience and Technology, 14th Annu Conf IEEE Eng Med Biol Soc Lyon. France: 1992; 46–50.
8. Ahonen AI, Hämäläinen MS, Kajola MJ, Knuutila JET, Laine PP, Lounasmaa OV, Simola JT, Tesche CD, Vilkman VA. A 122-channel magnetometer covering the whole head. Proceedings of the Satellite Symposium on Neuroscience and Technology, 14th Annu Conf IEEE Eng Med Biol Soc Lyon. France: 1992; 16–20.
9. Clarke J. SQUIDs: theory and practice. In: Weinstock H, Ralston RW, editors. The New Superconducting Electronics. Dordrecht, Boston: Kluwer Academic; 1993. p 123–180.
10. Vrba J, Robinson SE. Signal processing in magnetoencephalography. Methods 2001;25:249–271.
11. Jaklevic R, Lambe RC, Silver AH, Mercereau JE. Quantum interference effects in Josephson tunneling. Phys Rev Lett 1964;12:159–160.
12. Jaycox JM, Ketchen MB. Planar coupling scheme for ultra low noise dc SQUIDs. IEEE Trans Magn 1981;MAG-17:400–403.
13. Ketchen MB, Jaycox JM. Ultra low noise tunnel junction dc SQUID with a tightly coupled planar input coil. Appl Phys Lett 1982;40:736–738.
14. Clarke J, Goubau WM, Ketchen MB. Tunnel junction DC SQUID: fabrication, operation, and performance. J Low Temp Phys 1976;25:99–144.
15. McKay J, Vrba J, Betts K, Burbank MB, Lee S, Mori K, Nonis D, Spear P, Uriel Y. Implementation of multichannel biomagnetic measurement system using DSP technology. Proc Can Conf Elect Comp Eng 1993; 1090–1093.
16. Robinson SE. A digital SQUID controller for unshielded biomagnetic measurement. In: Aine C, Okada Y, Stroink G, Swithenby S, Wood C, editors. Biomag96: Proc 10th Int Conf Biomag; 2000; New York: Springer; p 103–106.
17. Knuutial JET, Ahonen AI, Hamalainen MS, Kajola MJ, Petteri Laine P, Lounasmaa OV, Parkkonen LT, Simola JTA, Tesche CD. A 122-channel whole-cortex SQUID system for measureing the brain's magnetic fields. IEEE Trans Mag 1993;29:3315–3321.
18. Vrba J. Multichannel SQUID biomagnetic systems. In: Weinstock H, editor. Applications of Superconductivity. Dordrecht: Kluwer Academic Publishers; 2000. p 61–138.
19. Kominis IK, Kornack TW, Allred JC, Romalis MV. A subfemtotesla multichannel atomic magnetometer. Nature (London) 2003;422:596–599.
20. Zimmerman JE. SQUID instruments and shielding for low level magnetic measurements. J Appl Phys 1977;48:702–710.
21. H. G. Vacuumschmelze GmbH, Shielded room model AK-3.
22. Sullivan GW, Flynn ER. Performance of the Los Alamos Shielded Room. In: Atsumi K, Kotani M, Ueno S, Katila T, Williamson SJ, editors. Biomagnetism '87. Tokyo: Tokyo Denki University Press; 1987. p 486–489.
23. Erné SN, Hahlbohm HD, Scheer G, Trontelj Z. The Berlin magnetically shielded room (BMSR). In: Erné SN, editor. Biomagnetism. Berlin: Walter de Gruyter; 1981. p 79–87.
24. Bork J, Hahlbohm HD, Klein R, Schnabel A. The 8-layered magnetically shielded room of the PTB: Design and construction. In: Nenonen J, Ilmoniemi R, Katila T, editors. Biomag 2000. Proc 12th Int Conf Biomag. Espoo, Finland: Helsinki University of Technology; 2001. p 970–973.
25. Matsuba H, Shintomi K, Yahara A, Irisawa D, Imai K, Yoshida H, Seike S. Superconducting shielding enclosing a human body for biomagnetic measurement. In: Baumgartner C, Deecke L, Stroink G, Williamson SJ, editors. Biomagnetism: Fundamental research and clinical applications. Amsterdam, The Netherlands: Elsevier Science, IOS Press; 1995. p 483–489.
26. Matsumoto K, Yamagishi Y, Wakusawa A, Noda T, Fujioka K, Kuraoka Y. SQUID based active shield for biomagnetic measurements. In: Hoke M, Erné SN, Okada Y, Romani Gl, editors. Biomagnetism: clinical aspects. Proc 8th Int Conf Biomag. Amsterdam: Excerpta Medica; 1992. p 857–861.
27. Malmivuo J, Lekkala J, Kontro P, Suomaa I, Vihinin H. Improvement of the properties of an eddy current magnetic shield with active compensation. J Phys E: Sci Instr 1987;20:151–164.
28. ter Brake HJM, Wieringa HJ, Rogalla H. Improvement of the performance of a μ-metal magnetically shielded room by means of active compensation. Meas Sci Technol 1991;2: 596–601.
29. Vrba J. SQUID gradiometers in real environments. In: Weinstock H, editor. SQUID sensors: Fundamentals, Fabrication, and Applications. Dordrecht, Boston: Kluwer Academic Publishers; 1996. p 117–178.
30. Vrba J, Robinson SE. SQUID sensor array configurations for magnetoencephalography applications. Supercond Sci Technol 2002;15:R51–R89.
31. Vrba J. Magnetoencephalography: the art of finding a needle in a haystack. Phys C 2002;368:1–9.
32. Huotilainen M, Ilmoniemi RJ, Tiitinen H, Lavikainen J, Alho K, Kajola M, Naatanen R. The projection method in removing eye blink artefacts from multichannel MEG measurements. In: Baumgartner C, Deecke L, Stroink G, Williamson SJ, editors. Biomagnetism: Fundamental research and clinical applications. Elsevier Science, IOS Press; 1995. p 363–367.
33. Tesche CD, Uusitalo MA, Ilmoniemi RJ, Huotilainen M, Kajola M, Salonen O. Signal-space projections of MEG data characterize both distributed and well-localized neuronal sources. Electroencephalogr Clin Neurophys 1995; 95:189–200.

34. Uusitalo MA, Ilmoniemi RJ. Signal-space projection method for separating MEG or EEG into components. Med Biol Eng Comput 1997;35:135–140.
35. Parkkonen LT, Simola JT, Tuoriniemi JT, Ahonen AI. An interference suppression system for multichannel magnetic field detector arrays. In: Yoshimoto T, Kotani M, Kuriki S, Karibe H, Nakasato N, editors. Recent Advances in Biomagnetism. Sendai: Tohoku University Press; 1999. p 13–16.
36. Taula S, Kajola MJ, Simola JT. The Signal Space Separation method. 14th Conf Int Soc Brain Electromagnetic Topography. Santa Fe, NM; 2003.
37. Ioannides AA, Mütter J, Barna-Popescu EA. Irreducible tensor representation of MEG signals: Theory and applications. In: Nenonen J, Ilmoniemi R, Katila T, editors. Biomag 2000. Proc 12th Int Conf Biomag. Espoo, Finland: Helsinki University of Technology; 2001. p 883–886.
38. ter Brake HJM. Cryogenic systems for superconducting devices. In: Weinstock H, editor. Applications of superconductivity. Dordrecht: Kluwer Academic Publishers; 2000. p 561–639.
39. VSM MedTech Ltd. (CTF) 9 Burbidge St. Coquitlam, B.C., Canada (www.vsmmedtech.com).
40. Nowak H. Biomagnetic Instrumentation. In: Andrä W, Nowak H, editors. Magnetism in Medicine. Berlin: Wiley VCH; 1998. p 88–135.
41. Hoenig HE, Daalmans GM, Bär L, Bömmel F, Paulus A, Uhl D, Weisse HJ, Schneider S, Seifert H, Reichenberger H, Abraham-Fuchs K, Multichannel DC. SQUID sensor array for biomagnetic applications. IEEE Trans Magn 1991; 27:2777–2785.
42. Tsukada K, Kandori A, Miyashita T, Sasabuchi H, Suzuki H, Kondo S, Komiyama Y, Teshogawara K. A simplified superconducting quantum interference device system to analyze vector components of a cardiac magnetic field. Proc 20th Annu Int Conf IEEE/EMBS. Hong Kong; 1998. p 524–527.
43. Van Leeuwen P, Haupt C, Hoormann C, Hailer B, Mackert BM, Stroink G. A 67 channel biomagnetometer designed for cardiology and other applications. In: Yoshimoto T, Kotani M, Kuriki S, Karibe H, Nakasato N, editors. Recent Advances in Biomagnetism. Sendai: Tohoku University Press; 1999. p 89–92.
44. Erné SN, Pasquarelli A, Kammrath H, Della Penna S, Torquati K, Pizzella V, Rossi R, Granata C, Russo M. Argos 55 - the new MCG system in Ulm. In: Yoshimoto T, Kotani M, Kuriki S, Karibe H, Nakasato N, editors. Recent Advances in Biomagnetism. Sendai: Tohoku University Press; 1999. p 27–30.
45. Montonen J, Ahonen A, Hämäläinen M, Ilmoniemi R, Laine P, Nenonen J, Paavola M, Simelius K, Simola J, Katila T. Magnetocardiographic functional imaging studies in BioMag Laboratory. In: Aine CJ, Okada Y, Stroink G, Swithenby S, Wood C, editors. Biomag 96: Proc Tenth Int Conf Biomag. New York: Springer; 2000. p 494–497.
46. ter Brake HJM, Janssen N, Flokstra J, Veldhuis D, Rogalla H. Multichannel heart scanner based on high-Tc SQUIDs. IEEE Trans Appl Supercond 1997;7:2545–2548.
47. Seidel P, Schmidl F, Wunderlich S, Dörrer L, Vogt T, Schneidewind H, Weidl R, Lösche R, Leder U, Solbig O, Nowak H. High-Tc SQUID systems for practical use. IEEE Trans Appl Supercond 1999;9:4077–4080.
48. Kouznesov KA, Borgmann J, Clarke CJS. High-Tc second-order gradiometer for magnetocardiography in an unshielded environment. Appl Phys Lett 1999;75:1979–1981.
49. Zhang Y, Panaitov G, Wang SG, Wolters N, Otto R, Schubert J, Zander W, Krause HJ, Soltner H, Bousack H, Braginski A. Second-order, high-temperature superconducting gradiometer for magnetocardiography in an unshielded environment. Appl Phys Lett 2000;76:906–908.
50. Ludwig F, Jansman ABM, Drung D, Lindström M, Bechstien S, Beyer J, Flokstra J, Schurig T. Optimization of direct-coupled high-Tc SQUID magnetometers for operation in a magnetically shielded environment. IEEE Trans Appl Supercond 2001;11:1824–1827.
51. Barthelmess H-J, Halverscheid M, Schiefenhövel B, Heim E, Schilling M, Zimmerman R. Low-noise biomagnetic measurements with a multichannel dc-SQUID system at 77 K. IEEE Trans Appl Supercond 2001;11:657–660.
52. Robinson SE, BM, FA, HG, KP, I Sekachev, Taylor B, Tillotson M, Wong VJG, Lowery C, Eswaran H, Wilson D, Murphy P, Preissl H. A biomagnetic instrument for human reproductive assessment. In: Nenonen J, Ilmoniemi R, Katila T, editors. Biomag2000, Proc 12th Int Conf Biomag. Espoo, Finland: Helsinki University of Technology; 2001. p 919–922.
53. 4-D Neuroimaging Inc. 9727 Pacific Heights Blvd., San Diego, CA 92121-3719 (www.4dneuroimaging.com).
54. Neuromag Oy, P.O. Box 68, FIN-00511 Helsinki, Finland (www.neuromag.com).
55. Eagle Technology, Inc. 1-2-23 Hirosaka, Kanazawa Ishikawa 920-0962, Japan (www.eagle-tek.com).
56. Advanced Technologies Biomagnetics S.r.l, Via Martiri di Pietransieri 2, 65129 Pescara, Italy (www.atb-it.com).
57. Polhemus Inc., Hercules Drive, P.O. Box 560, Colchester, VT 05446.
58. Bramidis PD, Ioannides AA. Combination of point and surface matching techniques for accurate registration of MEG and MRI. In: Aine CJ, Okada Y, Stroink G, Swithenby S, Wood C, editors. Biomag 96: Proc 10th Int Conf Biomag. New York: Springer; 1997. p 1126–1129.
59. Abraham-Fuchs K, Lindner L, Weganer P, Nestel F, Schneider S. Fusion of biomagnetism with MRI or CT images by contour fitting. Biomed Eng 1991;36(Suppl.): 88–89.
60. Kober H, Nimsky C, Vieth J, Fahlbusch R, Ganslandt O. Coregistration of function and anatomy in frameless stereotaxy by contour fitting, Stereotact Funct. Neurosurg 2002;79:272–283.
61. Braginski A, Krause HJ, Vrba J. SQUID magnetometers. In: Francombe MH, editor. Handbook of Thin Film Devices, Volume 3: Superconducting Film Devices. San Diego: Academic Press; 2000. p 149–225.
62. Farrell DE, Tripp JH, Zanzucchi PE. Magnetic measurements of human iron stores. IEEE Transactions on Magnetics May-1980;16:818–823.
63. Paulson DN, Fagaly RL, Toussaint RM, Fischer R. Biomagnetic susceptometer with SQUID instrumentation. IEEE Trans Magn 1991;27:3249–3252.
64. Fischer F. Liver iron susceptometry. In: Andrä W, Nowak H, editors. Magnetism in medicine: a handbook. Berlin; New York: Wiley-VCH; 1998. p 286–301.
65. Helmholtz H. Über einige Gesetze der Vertheilung elektrischer Ströme in köperlichen Leitern mit Anwendung auf die thierisch-elektrischen Versuche. Ann Phys Chem 1853;89:211–233, 353–377.
66. Hämäläinen M, Hari R, Ilmoniemi RJ, Knuutila JET, Lounasmaa OV. Magnetoencephalography-theory, instrumentation, and applications to noninvasive studies of the working human brain. Rev Mod Phys 1993;65:413–497.
67. Jeffs B, Leahy R, Singh M. An evaluation of methods for neuromagnetic image reconstruction. IEEE Trans Biomed Eng 1987;34:713–723.

68. Baillet S, Mosher JC, Leahy R. Electromagnetic brain mapping. IEEE Signal Proc Mag 2001;18:14–30.
69. Mosher JC, Leahy RM, Lewis PS. EEG and MEG: forward solutions for inverse methods. IEEE Trans Biomed Eng 1999;46:245–259.
70. Wikswo JP, Barach JP, Freeman JA. Magnetic field of a nerve impulse: first measurements. Science 1980;208:53–55.
71. Swinney KR, Wikswo, Jr. JP. A calculation of the magnetic field of a nerve action potential. Biophys J 1980;32:719–731.
72. Tesche CD. Non-invasive imaging of neuronal population dynamics in human thalamus. Brain Res 1996;729:253–258.
73. Tesche CD, Karhu J, Tissari SO. Non-invasive detection of neuronal population activity in human hippocampus. Brain Res Cogn Brain Res 1996;4:39–47.
74. Humphrey DR. Re-analysis of the antidromic cortical response. II. On the contribution of cell discharge and PSPs to the evoked potentials. Electroencephalogr Clin Neurophysiol 1968;25:421–442.
75. Creutzfeldt OD, Watanabe S, Lux HD. Relations between EEG phenomena and potentials of single cortical cells. I. Evoked responses after thalamic and epicortical stimulation. Electroencephalogr Clin Neurophysiol 1966;20:1–18.
76. Mitzdorf U. Current source-density method and application in cat cerebral cortex: investigation of evoked potentials and EEG phenomena. Physiol Rev 1985;65:37–100.
77. Okada YC, Wu J, Kyuhou S. Genesis of MEG signals in a mammalian CNS structure. Electroencephalogr Clin Neurophysiol 1997;103:474–485.
78. Okada Y. Toward understanding the physiological origins of neuromagnetic signals. In: Lu Z-L, Kaufman L, editors. Magnetic Source Imaging of the Brain. Mahwah. New Jersey: Lawrence Erlbaum Associates; 2003. p 43–76.
79. Weinberg H, Brickett P, Coolsma F, Baff M. Magnetic localisation of intracranial dipoles: simulation with a physical model. Electroencephalogr Clin Neurophysiol 1986;64:159–170.
80. Cheyne D, Roberts LE, Gaetz W, Bosnyak D, Nahmias C, Christoforou N, Weinberg H. Somatotopic organization of human somatosensory cortex: a comparison of EEG, MEG and fMRI methods. In: Koga Y, Nagata K, Hirata K, editors. Brain Topography Today. Amsterdam: Elsevier Science; 1998. p 76–81.
81. Sarvas J. Basic mathematical and electromagnetic concepts of the biomagnetic inverse problem. Phys Med Biol 1987;32:11–22.
82. Hillebrand A, Barnes GR. A quantitative assessment of the sensitivity of whole-head MEG to activity in the adult human cortex. Neuroimage 2002;16:638–650.
83. Lu ZL, Williamson SJ. Spatial extent of coherent sensory-evoked cortical activity. Exp Brain Res 1991;84:411–416.
84. Hämäläinen MS, Ilmoniemi RJ. Interpreting measured magnetic fields of the brain: Estimates of current distribution. Report TKK-F-A559. Helsinki University of Technology: Espoo, Finland, 1984.
85. Scherg M, Von Cramon D. Two bilateral sources of the late AEP as identified by a spatio-temporal dipole model. Electroencephalogr Clin Neurophysiol 1985;62:32–44.
86. Huang MX, Mosher JC, Leahy RM. A sensor-weighted overlapping-sphere head model and exhaustive head model comparison for MEG. Phys Med Biol 1999;44:423–440.
87. Gorodnitsky IF, George JS, Rao BD. Neuromagnetic source imaging with FOCUSS: a recursive weighted minimum norm algorithm. Electroencephalogr Clin Neurophysiol 1995;95:231–251.
88. Wagner M, Wischmann HA, Fuchs M, Kohler T, Drenckhahn R. Current density reconstruction using the L1 norm. In: Aine CJ, Okada Y, Stroink G, Swithenby S, Wood C, editors. Biomag96. New York: Springer-Verlag; 2000. p 393–396.
89. Ioannides AA, Bolton JPR, Clarke CJS. Continuous probabilistic solutions to the biomagnetic inverse problem. Inverse Problems 1990;6:523–542.
90. Pascual-Marqui RD, Michel CM, Lehmann D. Low resolution electromagnetic tomography: a new method for localizing electrical activity in the brain. Int J Psychophysiol 1994;18:49–65.
91. David O, Garnero L, Cosmelli D, Varela FJ. Estimation of neural dynamics from MEG/EEG cortical current density maps: application to the reconstruction of large-scale cortical synchrony. IEEE Trans Biomed Eng 2002;49: 975–987.
92. Baillet S, Garnero L. A Bayesian approach to introducing anatomo-functional priors in the EEG/MEG inverse problem. IEEE Trans Biomed Eng 1997;44:374–385.
93. Schmidt DM, George JS, Wood CC. Bayesian inference applied to the electromagnetic inverse problem. Hum Brain Mapp 1999;7:195–212.
94. Hämäläinen MS, Haario H, Lehtinen MS. Inferences about sources of neuromagnetic fields using Bayesian parameter estimation. Espoo, Finland: Helsinki University of Technolgy; 1987.
95. Sorenson HW. Parameter estimation. New York: Marcel Dekker; 1980.
96. Robinson SE, Rose DF. Current source estimation by spatially filtered MEG. In: HokeM, Erné SN, Okada Y, Romani GL, editors. Biomagnetism: clinical aspects. Proc 8th Int Conf Biomag. Amsterdam: Excerpta Medica; 1992. p 761–765.
97. Frost OL. An algorithm for linearly constrained adaptive array processing. Proc IEEE 1972;60:926–935.
98. Van Veen B, Buckley K. Beamforming: A versatile approach to spatial filtering, in IEEE ASSP Mag.; 1988. p 4–24.
99. Godara LC. Application of antenna array to mobile communications, Part II: Beam-Forming and direction-of-arrival considerations. Proc IEEE 1997;85:1195–1245.
100. Van Veen BD, Van Drongelen W, Yuchtman M, Suzuki A. Localization of brain electrical activity via linearly constrained minimum variance spatial filtering. IEEE Trans Biomed Eng 1997;44:867–880.
101. Robinson SE, Vrba J. Functional neuroimaging by synthetic aperture magnetometry (SAM). In: Yoshimoto T, Kotani M, Kuriki S, Karibe H, Nakasato N, editors. Recent Advances in Biomagnetism. Sendai: Tohoku University Press; 1999. p 302–305.
102. Vrba J, Robinson SE. Linearly constrained minimum variance beamformers, synthetic aperture magnetometry and MUSIC in MEG applications. IEEE, Proc 34th Asilomar Conf. Signals, Systems, Comput. Pacific Grove, CA: Omnipress; 2000. p 313–317.
103. Schmidt RO. Multiple emitter location and signal parameter estimation. IEEE Trans Anten Propagat 1986;AP-34:276–280.
104. Mosher JC, Lewis PS, Leahy RM. Multiple dipole modeling and localization from spatio-temporal MEG data. IEEE Trans Biomed Eng 1992;39:541–557.
105. Mosher JC, Leahy R. Source localization using recursively applied and projected (RAP) MUSIC. IEEE Trans Signal Proc 1999;47:332–340.
106. Maier J, Dagnelie G, Spekreijse H, van Dijk BW. Principal components analysis for source localization of VEPs in man. Vision Res 1987;27:165–177.
107. Achim A, Richer F, Saint-Hilaire JM. Methods for separating temporally overlapping sources of neuroelectric data. Brain Topogr 1988;1:22–28.

108. Jung TP, Makeig S, Mckeown MJ, Bell AJ, L. T, Sejnowski TJ. Imaging brain dynamics using independent component analysis. Proc IEEE 2001;89:1107–1122.
109. Cardoso J-F. Blind signal separation: Statistical principles. Proc IEEE 1998;86:2009–2025.
110. Hyvarinen A. Fast and robust fixed-point algorithms for independent component analysis. IEEE Trans Neural Networks 1999;10:626–634.
111. Makeig S, Jung TP, Bell AJ, Ghahremani D, Sejnowski TJ. Blind separation of auditory event-related brain responses into independent components. Proc Natl Acad Sci USA 1997;94:10979–10984.
112. Ziehe A, Muller KR, Nolte G, Mackert BM, Curio G. Artifact reduction in magnetoneurography based on time-delayed second-order correlations. IEEE Trans Biomed Eng 2000; 47:75–87.
113. Lewine JD, Orrison, Jr. WW. Magnetic source imaging: basic principles and applications in neuroradiology. Acad Radiol 1995;2:436–440.
114. Wheless JW, Castillo E, Maggio V, Kim HL, Breier JI, Simos PG, Papanicolaou AC. Magnetoencephalography (MEG) and magnetic source imaging (MSI). Neurologist 2004; 10:138–153.
115. Lu Z-L, Kaufman L, editors. Magnetic Source Imaging of the Brain. Mahwah, NJ: Lawrence Erlbaum Associates; 2003.
116. Gallen CC, Schwartz BJ, Bucholz RD, Malik G, Barkley GL, Smith J, Tung H, Copeland B, Bruno L, Assam S. Presurgical localization of functional cortex using magnetic source imaging. J Neurosurg 1995;82:988–994.
117. Roberts TP, Poeppel D, Rowley HA. Magnetoencephalography and magnetic source imaging. Neuropsychiat Neuropsychol Behav Neurol 1998;11:49–64.
118. Brenner D, Lipton J, Kaufman L, Williamson SJ. Somatically evoked magnetic fields of the human brain. Science 1978;199:81–83.
119. Kakigi R, Hoshiyama M, Shimojo M, Naka D, Yamasaki H, Watanabe S, Xiang J, Maeda K, Lam K, Itomi K, Nakamura A. The somatosensory evoked magnetic fields. Prog Neurobiol 2000;61:495–523.
120. Nakamura A, Yamada T, Goto A, Kato T, Ito K, Abe Y, Kachi T, Kakigi R. Somatosensory homunculus as drawn by MEG. Neuroimage 1998;7:377–386.
121. Curio G, Mackert BM, Burghoff M, Koetitz R, Abraham-Fuchs K, Harer W. Localization of evoked neuromagnetic 600 Hz activity in the cerebral somatosensory system. Electroencephalogr Clin Neurophysiol 1994;91:483–487.
122. Hashimoto I, Mashiko T, Imada T. Somatic evoked high-frequency magnetic oscillations reflect activity of inhibitory interneurons in the human somatosensory cortex. Electroencephalogr Clin Neurophysiol 1996;100:189–203.
123. Ikeda H, Leyba L, Bartolo A, Wang Y, Okada YC. Synchronized spikes of thalamocortical axonal terminals and cortical neurons are detectable outside the pig brain with MEG. J Neurophysiol 2002;87:626–630.
124. Hari R, Karhu J, Hämäläinen MS, Knuutila J, Salonen O, Sams M, Vilkman V. Functional organization of the human first and second somatosensory cortices: a neuromagnetic study. Eur J Neurosci 1993;5:724–734.
125. Suk J, Ribary U, Cappell J, Yamamoto T, Llinas R. Anatomical localization revealed by MEG recordings of the human somatosensory system. Electroencephalogr Clin Neurophysiol 1991;78:185–196.
126. Flor H, Elbert T, Knecht S, Wienbruch C, Pantev C, Birbaumer N, Larbig W, Taub E. Phantom-limb pain as a perceptual correlate of cortical reorganization following arm amputation. Nature (London) 1995;375:482–484.
127. Elbert T, Pantev C, Wienbruch C, Rockstroh B, Taub E. Increased cortical representation of the fingers of the left hand in string players. Science 1995;270:305–307.
128. Deecke L, Weinberg H, Brickett P. Magnetic fields of the human brain accompanying voluntary movement: Bereitschaftsmagnetfeld. Exp Brain Res 1982;48:144–148.
129. Cheyne D, Weinberg H. Neuromagnetic fields accompanying unilateral finger movements: pre-movement and movement-evoked fields. Exp Brain Res 1989;78:604–612.
130. Kristeva R, Cheyne D, Lang W, Lindinger G, Deecke L. Movement-related potentials accompanying unilateral and bilateral finger movements with different inertial loads. Electroencephalogr Clin Neurophysiol 1990;75:410–418.
131. Cheyne D, Endo H, Takeda T, Weinberg H. Sensory feedback contributes to early movement-evoked fields during voluntary finger movements in humans. Brain Res 1997;771:196–202.
132. Kelso JA, Fuchs A, Lancaster R, Holroyd T, Cheyne D, Weinberg H. Dynamic cortical activity in the human brain reveals motor equivalence. Nature (London) 1998;392:814–818.
133. Endo H, Kizuka T, Masuda T, Takeda T. Automatic activation in the human primary motor cortex synchronized with movement preparation. Cogn Brain Res 1999;8:229–239.
134. Xiang J, Hoshiyama M, Koyama S, Kaneoke Y, Suzuki H, Watanabe S, Naka D, Kakigi R. Somatosensory evoked magnetic fields following passive finger movement. Brain Res Cogn Brain Res 1997;6:73–82.
135. Lang W, Cheyne D, Hollinger P, Gerschlager W, Lindinger G. Electric and magnetic fields of the brain accompanying internal simulation of movement. Cogn Brain Res 1996;3:125–129.
136. Conway BA, Halliday DM, Farmer SF, Shahani U, Maas P, Weir AI, Rosenberg JR. Synchronization between motor cortex and spinal motoneuronal pool during the performance of a maintained motor task in man. J Physiol 1995;489(Pt. 3): 917–924.
137. Brown P. Cortical drives to human muscle: the Piper and related rhythms. Prog Neurobiol 2000;60:97–108.
138. Salenius S, Avikainen S, Kaakkola S, Hari R, Brown P. Defective cortical drive to muscle in Parkinson's disease and its improvement with levodopa. Brain 2002;125: 491–500.
139. Timmermann L, Gross J, Dirks M, Volkmann J, Freund HJ, Schnitzler A. The cerebral oscillatory network of parkinsonian resting tremor. Brain 2003;126:199–212.
140. Hari R, Salmelin R. Human cortical oscillations: a neuromagnetic view through the skull. Trends Neurosci 1997;20: 44–49.
141. Cheyne D, Gaetz W, Garnero L, Lachaux JP, Ducorps A, Schwartz D, Varela FJ. Neuromagnetic imaging of cortical oscillations accompanying tactile stimulation. Brain Res Cogn Brain Res 2003;17:599–611.
142. Feige B, Kristeva-Feige R, Rossi S, Pizzella V, Rossini PM. Neuromagnetic study of movement-related changes in rhythmic brain activity. Brain Res 1996;734:252–260.
143. Gaetz WC, Cheyne DO. Localization of human somatosensory cortex using spatially filtered magnetoencephalography. Neurosci Lett 2003;340:161–164.
144. Taniguchi M, Kato A, Fujita N, Hirata M, Tanaka H, Kihara T, Ninomiya H, Hirabuki N, Nakamura H, Robinson SE, Cheyne D, Yoshimine T. Movement-related desynchronization of the cerebral cortex studied with spatially filtered magnetoencephalography. Neuroimage 2000;12:298–306.
145. Brenner D, Williamson SJ, Kaufman L. Visually evoked magnetic fields of the human brain. Science 1975;190:480–482.
146. Supek S, Aine CJ, Ranken D, Best E, Flynn ER, Wood CC. Single vs. paired visual stimulation: superposition of early

neuromagnetic responses and retinotopy in extrastriate cortex in humans. Brain Res 1999;830:43–55.
147. Fylan F, Holliday IE, Singh KD, Anderson SJ, Harding GF. Magnetoencephalographic investigation of human cortical area V1 using color stimuli. Neuroimage 1997;6:47–57.
148. Singh KD, Barnes GR, Hillebrand A, Forde EM, Williams AL. Task-related changes in cortical synchronization are spatially coincident with the hemodynamic response. Neuroimage 2002;16:103–114.
149. Anderson SJ, Holliday IE, Singh KD, Harding GF. Localization and functional analysis of human cortical area V5 using magneto-encephalography. Proc R Soc London Sev B Biol Sci 1996;263:423–431.
150. Maruyama K, Kaneoke Y, Watanabe K, Kakigi R. Human cortical responses to coherent and incoherent motion as measured by magnetoencephalography. Neurosci Res 2002;44:195–205.
151. Anderson SJ, Holliday IE, Harding GF. Assessment of cortical dysfunction in human strabismic amblyopia using magnetoencephalography (MEG). Vision Res 1999;39:1723–1738.
152. Reite M, Edrich J, Zimmerman JT, Zimmerman JE. Human magnetic auditory evoked fields. Electroencephalogr Clin Neurophysiol 1978;45:114–117.
153. Romani GL, Williamson SJ, Kaufman L, Brenner D. Characterization of the human auditory cortex by the neuromagnetic method. Exp Brain Res 1982;47:381–393.
154. Pantev C, Hoke M, Lehnertz K, Lutkenhoner B, Anogianakis G, Wittkowski W. Tonotopic organization of the human auditory cortex revealed by transient auditory evoked magnetic fields. Electroencephalogr Clin Neurophysiol 1988;69: 160–170.
155. Pantev C, Lutkenhoner B. Magnetoencephalographic studies of functional organization and plasticity of the human auditory cortex. J Clin Neurophysiol 2000;17:130–142.
156. Roberts TP, Ferrari P, Stufflebeam SM, Poeppel D. Latency of the auditory evoked neuromagnetic field components: stimulus dependence and insights toward perception. J Clin Neurophysiol 2000;17:114–129.
157. Weinberg H, Cheyne D, Brickett P, Harrop R, Gordon R. An interaction of cortical sources associated with simultaneous auditory and somatosensory stimulation. In: Pfurtscheller G, Lopes da Silva FH, editors. Functional Brain Imaging. Lewiston, N.Y.: Hans Huber Publishers; 1988. p 83–88.
158. Ribary U, Ioannides AA, Singh KD, Hasson R, Bolton JP, Lado F, Mogilner A, Llinas R. Magnetic field tomography of coherent thalamocortical 40-Hz oscillations in humans. Proc Natl Acad Sci USA 1991;88:11037–11041.
159. Ross B, Picton TW, Pantev C. Temporal integration in the human auditory cortex as represented by the development of the steady-state magnetic field. Hear Res 2002;165:68–84.
160. Starr A, Kristeva R, Cheyne D, Lindinger G, Deecke L. Localization of brain activity during auditory verbal short-term memory derived from magnetic recordings. Brain Res 1991;558:181–190.
161. Mecklinger A, Maess B, Opitz B, Pfeifer E, Cheyne D, Weinberg H. A MEG analysis of the P300 in visual discrimination tasks. Electroencephalogr Clin Neurophysiol 1998;108:45–56.
162. Alho K. Cerebral generators of mismatch negativity (MMN) and its magnetic counterpart (MMNm) elicited by sound changes. Ear Hear 1995;16:38–51.
163. Halgren E, Raij T, Marinkovic K, Jousmaki V, Hari R. Cognitive response profile of the human fusiform face area as determined by MEG. Cereb Cortex 2000;10:69–81.
164. Salmelin R, Hari R, Lounasmaa OV, Sams M. Dynamics of brain activation during picture naming. Nature (London) 1994;368:463–465.
165. Sams M, Aulanko R, Hamalainen M, Hari R, Lounasmaa OV, Lu ST, Simola J. Seeing speech: visual information from lip movements modifies activity in the human auditory cortex. Neurosci Lett 1991;127:141–145.
166. Pulvermuller F, Shtyrov Y, Ilmoniemi R. Spatiotemporal dynamics of neural language processing: an MEG study using minimum-norm current estimates. Neuroimage 2003;20:1020–1025.
167. Pylkkanen L, Marantz A. Tracking the time course of word recognition with MEG. Trends Cogn Sci 2003;7:187–189.
168. Roberts TP, Ferrari P, Poeppel D. Latency of evoked neuromagnetic M100 reflects perceptual and acoustic stimulus attributes, Neuroreport 1998;9:3265–3269.
169. Helenius P, Salmelin R, Service E, Connolly JF. Semantic cortical activation in dyslexic readers. J Cogn Neurosci 1999;11:535–550.
170. Salmelin R, Schnitzler A, Schmitz F, Jancke L, Witte OW, Freund HJ. Functional organization of the auditory cortex is different in stutterers and fluent speakers. Neuroreport 1998;9:2225–2229.
171. Tallon-Baudry C, Bertrand O. Oscillatory gamma activity in humans and its role in object representation. Trends Cogn Sci 1999;3:151–162.
172. Nishitani N, Hari R. Temporal dynamics of cortical representation for action. Proc Natl Acad Sci USA 2000;97:913–918.
173. Varela F, Lachaux JP, Rodriguez E, Martinerie J. The brainweb: phase synchronization and large-scale integration. Nat Rev Neurosci 2001;2:229–239.
174. Gallen CC, Sobel DF, Waltz T, Aung M, Copeland B, Schwartz BJ, Hirschkoff EC, Bloom FE. Noninvasive presurgical neuromagnetic mapping of somatosensory cortex. Neurosurgery 1993;33:260–268; discussion 268.
175. Papanicolaou AC, Simos PG, Castillo EM, Breier JI, Sarkari S, Pataraia E, Billingsley RL, Buchanan S, Wheless J, Maggio V, Maggio WW. Magnetocephalography: a noninvasive alternative to the Wada procedure. J Neurosurg 2004;100:867–876.
176. Naatanen R, Lehtokoski A, Lennes M, Cheour M, Huotilainen M, Iivonen A, Vainio M, Alku P, Ilmoniemi RJ, Luuk A, Allik J, Sinkkonen J, Alho K. Language-specific phoneme representations revealed by electric and magnetic brain responses. Nature (London) 1997;385:432–434.
177. Bowyer SM, Moran JE, Mason KM, Constantinou JE, Smith BJ, Barkley GL, Tepley N. MEG localization of language-specific cortex utilizing MR-FOCUSS. Neurology 2004;62: 2247–2255.
178. Firsching R, Bondar I, Heinze HJ, Hinrichs H, Hagner T, Heinrich J, Belau A. Practicability of magnetoencephalography-guided neuronavigation. Neurosurg Rev 2002; 25:73–78.
179. Ganslandt O, Fahlbusch R, Nimsky C, Kober H, Moller M, Steinmeier R, Romstock J, Vieth J. Functional neuronavigation with magnetoencephalography: outcome in 50 patients with lesions around the motor cortex. J Neurosurg 1999;91:73–79.
180. Rosenow F, Luders H. Presurgical evaluation of epilepsy. Brain 2001;124:1683–1700.
181. Stefan H, Hummel C, Scheler G, Genow A, Druschky K, Tilz C, Kaltenhauser M, Hopfengartner R, Buchfelder M, Romstock J. Magnetic brain source imaging of focal epileptic activity: a synopsis of 455 cases. Brain 2003.
182. Barkley GL. Controversies in neurophysiology. MEG is superior to EEG in localization of interictal epileptiform activity. Pro Clin Neurophysiol 2004;115:1001–1009.
183. Baumgartner C. Controversies in clinical neurophysiology. MEG is superior to EEG in the localization of interictal epileptiform activity: Con. Clin Neurophysiol 2004;115: 1010–1020.

184. Otsubo H, Snead III OC. Magnetoencephalography and magnetic source imaging in children. J Child Neurol 2001;16:227–235.
185. Stefan H, Hummel C, Hopfengartner R, Pauli E, Tilz C, Ganslandt O, Kober H, Moler A, Buchfelder M. Magnetoencephalography in extratemporal epilepsy. J Clin Neurophysiol 2000;17:190–200.
186. Baumgartner C, Pataraia E, Lindinger G, Deecke L. Magnetoencephalography in focal epilepsy. Epilepsia 2000;41 (Suppl 3): S39–47.
187. de Jongh A, Baayen JC, de Munck JC, Heethaar RM, Vandertop WP, Stam CJ. The influence of brain tumor treatment on pathological delta activity in MEG. Neuroimage 2003;20:2291–2301.
188. Taniguchi M, Kato A, Ninomiya H, Hirata M, Cheyne D, Robinson SE, Maruno M, Saitoh Y, Kishima H, Yoshimine T. Cerebral motor control in patients with gliomas around the central sulcus studied with spatially filtered magnetoencephalography. J Neurol Neurosurg Psychiat 2004;75: 466–471.
189. Schiffbauer H, Ferrari P, Rowley HA, Berger MS, Roberts TP. Functional activity within brain tumors: a magnetic source imaging study. Neurosurgery 2001;49:1313–1320; discussion 1320–1311.
190. Gallen CC, Tecoma E, Iragui V, Sobel DF, Schwartz BJ, Bloom FE. Magnetic source imaging of abnormal low-frequency magnetic activity in presurgical evaluations of epilepsy. Epilepsia 1997;38:452–460.
191. Rossini PM, Tecchio F, Pizzella V, Lupoi D, Cassetta E, Pasqualetti P, Paqualetti P. Interhemispheric differences of sensory hand areas after monohemispheric stroke: MEG/MRI integrative study. Neuroimage 2001;14:474–485.
192. Gallien P, Aghulon C, Durufle A, Petrilli S, De Crouy AC, Carsin M, Toulouse P. Magnetoencephalography in stroke: a 1-year follow-up study. Eur J Neurol 2003;10: 373–382.
193. Lewine JD, Davis JT, Sloan JH, Kodituwakku PW, Orrison Jr. WW. Neuromagnetic assessment of pathophysiologic brain activity induced by minor head trauma. Am J Neuroradiol 1999;20:857–866.
194. Bowyer SM, Aurora KS, Moran JE, Tepley N, Welch KM. Magnetoencephalographic fields from patients with spontaneous and induced migraine aura. Ann Neurol 2001;50:582–587.
195. Tran TD, Inui K, Hoshiyama M, Lam K, Qiu Y, Kakigi R. Cerebral activation by the signals ascending through unmyelinated C-fibers in humans: a magnetoencephalographic study. Neuroscience 2002;113:375–386.
196. Inui K, Tran TD, Qiu Y, Wang X, Hoshiyama M, Kakigi R. Pain-related magnetic fields evoked by intra-epidermal electrical stimulation in humans. Clin Neurophysiol 2002;113: 298–304.
197. Kamada K, Moller M, Saguer M, Ganslandt O, Kaltenhauser M, Kober H, Vieth J. A combined study of tumor-related brain lesions using MEG and proton MR spectroscopic imaging. J Neurol Sci 2001;186:13–21.
198. Reite M, Teale P, Rojas DC. Magnetoencephalography: applications in psychiatry. Biol Psychiat 1999;45:1553–1563.
199. Pekkonen E, Hirvonen J, Jaaskelainen IP, Kaakkola S, Huttunen J. Auditory sensory memory and the cholinergic system: implications for Alzheimer's disease. Neuroimage 2001;14:376–382.
200. Nishitani N, Avikainen S, Hari R. Abnormal imitation-related cortical activation sequences in Asperger's syndrome. Ann Neurol 2004;55:558–562.
201. Hren R, Zhang X, Stroink G. Comparison between electrocardiographic and magnetocardiographic inverse solutions using the boundary element method. Med Biol Eng Comput 1996;34:110–114.
202. Fenici R, Melillo G. Magnetocardiography: ventricular arrhythmias. Eur Heart J 1993;14(Suppl E): 53–60.
203. Stroink G, Moshage W, Achenbach S. Cardiomagnetism. In: Andrä W, Nowak H, editors. Magnetism in Medicine: A Handbook. Berlin: Wiley; 1998. p 136–189.
204. Tavarozzi I, Comani S, Del Gratta C, Di Luzio S, Romani GL, Gallina S, Zimarino M, Brisinda D, Fenici R, De Caterina R. Magnetocardiography: current status and perspectives. Part II: Clinical applications. Ital Heart J 2002;3:151–165.
205. Fenici R, Brisinda D, Nenonen J, Fenici P. Noninvasive study of ventricular preexcitation using multichannel magnetocardiography. Pacing Clin Electrophysiol 2003;26: 431–435.
206. Kanzaki H, Nakatani S, Kandori A, Tsukada K, Miyatake K. A new screening method to diagnose coronary artery disease using multichannel magnetocardiogram and simple exercise. Basic Res Cardiol 2003;98:124–132.
207. Leder U, Pohl HP, Michaelsen S, Fritschi T, Huck M, Eichhorn J, Muller S, Nowak H. Noninvasive biomagnetic imaging in coronary artery disease based on individual current density maps of the heart. Int J Cardiol 1998; 64:83–92.
208. Blum T, Saling E, Bauer R. First magnetoencephalographic recordings of the brain activity of a human fetus. Br J Obstet Gynaecol 1985;92:1224–1229.
209. Kariniemi V, Ahopelto J, Karp PJ, Katila TE. The fetal magnetocardiogram. J Perinat Med 1974;2:214–216.
210. Lowery CL, Campbell JQ, Wilson JD, Murphy P, Preissl H, Malak SF, Eswaran H. Noninvasive antepartum recording of fetal S-T segment with a newly developed 151-channel magnetic sensor system. Am J Obstet Gynecol 2003;188: 1491–1496; discussion 1496–1497.
211. Wakai RT, Leuthold AC, Martin CB. Atrial and ventricular fetal heart rate patterns in isolated congenital complete heart block detected by magnetocardiography. Am J Obstet Gynecol 1998;179:258–260.
212. Quartero HW, Stinstra JG, Golbach EG, Meijboom EJ, Peters MJ. Clinical implications of fetal magnetocardiography. Ultrasound Obstet Gynecol 2002;20:142–153.
213. Kahler C, Schleussner E, Grimm B, Schneider U, Haueisen J, Vogt L, Seewald HJ. Fetal magnetocardiography in the investigation of congenital heart defects. Early Hum Dev 2002;69:65–75.
214. Van Leeuwen P. Future topics in fetal magnetocardiography. In: Nenonen J, Ilmoniemi R, Katila T, editors. Biomag 2000: Proc 12th Int Conf Biomag. Espoo, Finland: Helsinki University of Technology; 2001. p 587–590.
215. Lengle JM, Chen M, Wakai RT. Improved neuromagnetic detection of fetal and neonatal auditory evoked responses. Clin Neurophysiol 2001;112:785–792.
216. Schneider U, Schleussner E, Haueisen J, Nowak H, Seewald HJ. Signal analysis of auditory evoked cortical fields in fetal magnetoencephalography. Brain Topogr 2001;14:69–80.
217. Schleussner E, Schneider U, Olbertz D, Kahler R, Huonker R, Michels W, Nowak H, Seewald HJ. Assessment of the fetal neuronal maturation using auditory evoked fields in fetal magnetoencephalography. In: Yoshimoto T, Kotani M, Kuriki S, Karibe H, Nakasato N, editors. Recent Advances in Biomagnetism. Sendai: Tohoku University Press; 1999. p 975–977.

218. Lowery C, Robinson S, Eswaran H, V. J, H. G, Cheung T. Detection of the transient and steady-state auditory evoked responses in the human fetus. In: Yoshimoto T, Kotani M, Kuriki S, Karibe H, Nakasato N, editors. Recent Advances in Biomagnetism. Sendai: Tohoku University Press; 1999. p 963–966.
219. Rose D, Eswaran H. Spontaneous neuronal activity in fetuses and newborns. Exp Neurol, in press.
220. Eswaran H, Wilson J, Preissl H, Robinson S, Vrba J, Murphy P, Rose D, Lowery C. Magnetoencephalographic recordings of visual evoked brain activity in the human fetus. Lancet 2002;360:779–780.
221. Vrba J, Robinson SE, McCubbin J, Murphy P, Eswaran H, Wilson JD, Preissl H, Lowery CL. Human fetal brain imaging by magnetoencephalography: verification of fetal brain signals by comparison with fetal brain models. Neuroimage 2004;21:1009–1020.
222. Paulson DN, Fagaly RL, Toussaint RM, Fischer F. Biomagnetic susceptometry with SQUID instrumentation. IEEE Trans Mag 1991;27:3249–3252.
223. Brittenham GM, Sheth S, Allen CJ, Farrell DE. Noninvasive methods for quantitative assessment of transfusional iron overload in sickle cell disease. Semin Hematol 2001; 38: 37–56.
224. Hoshiyama M, Kakigi R, Nagata O. Peripheral nerve conduction recorded by a micro gradiometer system (micro-SQUID) in humans. Neurosci Lett 1999;272:199–202.
225. Mackert BM, Curio G, Burghoff M, Trahms L, Marx P. Magnetoneurographic 3D localization of conduction blocks in patients with unilateral S1 root compression. Electroencephalogr Clin Neurophysiol 1998;109:315–320.
226. Le Gros V, Lemaigre D, Suon C, Pozzi JP, Liot F. Magnetopneumography: a general review. Eur Respir J 1989;2: 149–159.
227. Huvinen M, Oksanen L, Kalliomaki K, Kalliomaki PL, Moilanen M. Estimation of individual dust exposure by magnetopneumography in stainless steel production. Sci Total Environ 1997;199:133–139.
228. Moller W, Barth W, Kohlhaufl M, Haussinger K, Stahlhofen W, Heyder J. Human alveolar long-term clearance of ferromagnetic iron oxide microparticles in healthy and diseased subjects. Exp Lung Res 2001;27:547–568.
229. Moraes R, Toncon LE, Baffa O, Oba-Kunyioshi AS, Wakai RT, Leuthold AC. Adaptive, autoregressive spectral estimation for analysis of electrical signals of gastric origin. Physiol Meas 2003;24:91–106.
230. Kosch O, Osmanoglou E, Hartman V, Strenzke A, Weitschies W, Wiedenmann B, Monnikes H, Trahms L. Investigation of gastrointestinal transport by magnetic marker localization. Biomed Tech (Berl) 2002;47(Suppl 1, Pt 2): 506–509.

See also ELECTROCARDIOGRAPHY, COMPUTERS IN; ELECTROENCEPHALOGRAPHY; EVOKED POTENTIALS; PULMONARY PHYSIOLOGY.

BIOMATERIALS, ABSORBABLE

MARK BORDEN
Director of Biomaterials Research
Irvine, California

INTRODUCTION

Historically, the use of implants in orthopedic surgery has originated from fracture repair and joint replacement applications. During the late 1920s, stainless-steel bone implants such as Kirshner nails and Steinman pins were popularized for the surgical treatment of fractures (1). With the introduction of new surgical materials such as cobalt alloys, polyethylene and poly(tetrafluoroethylene) [Teflon], surgeons and engineers began working toward the design and fabrication of artificial joints. The advent of new high strength implant materials allowed researchers such Dr. John Charnley to begin pioneering work in total hip replacement surgery in the late 1930s (1,2). As advances in chemistry, metallurgy, and ceramics progressed throughout the years, a large variety of implants have entered the orthopedic market. Today, orthopedic implants are composed of specialized metals, ceramics, polymers, and composites that possess a large range in properties. Although these materials have been successfully fabricated into a variety of implants, one common issue has remained. Once the device has performed its required function and is no longer needed, it remains as a bystander in the now healthy tissue. The issue is that the long-term presence of an implant in the body can result in implant-related complications such as loosening, migration, mechanical breakdown and fatigue, generation of wear particles, and other negative effects (3–6). With prolonged patient life spans and higher activity levels, more and more people are now outliving the lifetime of their implants.

The potential for long-term implant problems has driven researchers to look to a unique category of materials that are capable of being completely resorbed by the body. These bioresorbable or biodegradable materials are characterized by the ability to be chemically broken down into harmless byproducts that are metabolized or excreted by the body. Materials of this type offer a great advantage over conventional nonresorbable implant materials. Bioresorbable implants provide their required function until the tissue is healed, and once their role is complete, the implant is completely resorbed by the body. The end result is healthy tissue with no signs that an implant was ever present. As the implant is completely gone from the site, long-term complications associated with nonresorbable devices do not exist.

ORTHOPEDIC APPLICATIONS OF RESORBABLE IMPLANTS

The ability of a resorbable implant to provide temporary fixation followed by complete resorption is a desirable property for a large variety of surgical applications. In relation to orthopedic surgery, this behavior is particularly useful based on the goal of restoring physiological function to the tissues and joints of the skeleton. In general, orthopedic surgery is often compared with carpentry in that the surgeon's instruments often consist of hammers, drills, and saws. Similar to carpentry, specialized screws, plates, pins, and nails are used to fix one material to another. In orthopedics, this fixation can be categorized into two main areas: bone-to-bone fixation and soft tissue-to-bone fixation. Bone fixation is used in the treatment of complex fractures and in reconstructive procedures of the skeleton. The implants used in these surgeries are designed to

maintain the position of the bone fragments, to stabilize the site, and to allow for eventual fusion of the fracture. As a result of the fracture healing process, the bone is remodeled so effectively that it is often difficult to locate the initial injury. With nonresorbable implants, the long-term presence of the device only serves as a source for potential complications. Resorbable implants, on the other hand, alleviate this concern by fully resorbing and allowing the bone to completely remodel into its normal physiological state.

In addition to bone fixation, soft tissue fixation is also an excellent application of resorbable implants. This type of reconstruction is often the result of trauma to joints such as the knee and shoulder. Typically developing from sports injuries or accidents, the goal is to restore stability to the joint by replacing or reconstructing the ligament or tendon interface to bone. In the knee, for example, the reconstruction of a torn anterior cruciate ligament (ACL) is a common sports medicine procedure. This type of surgical reconstruction consists of replacing the torn ACL with a bone-tendon-bone graft taken from the patient's patella and fixing the graft across the joint. During the procedure, the bony portion of the ACL graft is fixed in bone tunnels drilled into the tibia and femur. In order to stabilize the graft and aid in the formation of a stable bone-to-ligament interface, interference screws are used to fix the graft to the site. Once bone has been incorporated into the graft, the device is no longer needed.

Another example of soft tissue reconstruction is the repair of a tear in the rotator cuff tendon of the shoulder. This type of injury requires reestablishing the tendon-to-bone interface. To facilitate this process and restore stability to the shoulder, implants called suture anchors are used to provide a means to affix the torn tendon to the bone of the humerous. Just as the name describes, these implants function by providing an anchor in bone that allows the attached suture to tighten down on the tendon and pull it in contact with bone. As healing progresses, a stable interface develops and joint function is restored. Similar to other fixation applications, once the interface has fully healed, the implant is no longer needed.

FUNCTION OF A RESORBABLE IMPLANT

As seen from the various types of tissue fixation procedures within orthopedic surgery, resorbable implants are exposed to a variety of healing environments. Out of the currently used materials in orthopedic surgery, only the polymer and ceramic groups contain resorbable biomaterials. It is the specific properties of these materials that allow them to be used as resorbable devices. In evaluating a material for potential use as an implant, the key properties include implant biocompatibility, resorbability, and mechanical properties. The first criteria, biocompatibility, refers to the ability of the material to be implanted into the body without negatively affecting the surrounding tissue, which includes the absence of inflammation, toxicity (materials that kill surrounding cells), carcinogenicity (materials that can cause cancer), genotoxicicty (materials that damage the DNA of surrounding cells), and mutagenicity (materials that cause genetic mutations within the cell). More specifically related to bone, the implant must also be osteocompatible, which means the material does not interfere with the normal bone healing process (7).

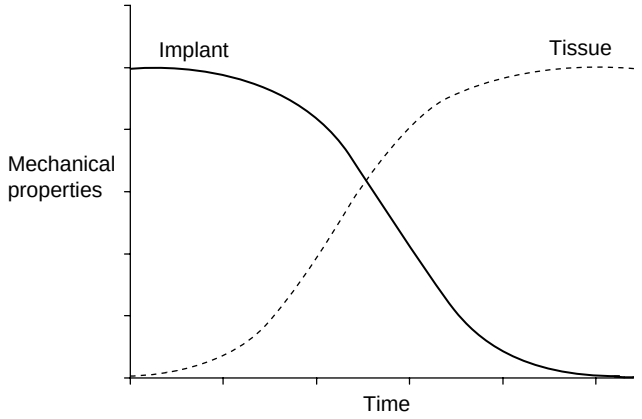


Figure 1. Optimal stress transfer profile for resorbable implants demonstrating the load-sharing properties of the implant.

Although biocompatibility has a direct effect on how the tissue surrounding the device heals and is an important property of the implant, the main criteria related to implant function are the resorbability and mechanical properties. Once the device is implanted, it provides immediate mechanical stabilization to the site while the tissue heals. As the regenerating bone, ligaments, or tendons become stronger over time, the implant site becomes less dependent on the device and more dependent on the healing tissue. This concept is shown in Fig. 1. In this situation, the implant provides all of the mechanical support immediately following placement. As the device begins to degrade, the mechanical properties decrease over time and are gradually transferred to the new tissue. During this period, the regenerating tissue responds to the gradual loads and begins to remodel and become stronger. In the healing of musculoskeletal tissue, the sharing of the load between the implant and the tissue results in further regeneration. Once healing is complete, the load is fully transitioned to the tissue, which is now mechanically independent from the implant. Upon final resorption of the device, the site is left fully functional and entirely free of any implant material.

The ability to gradually transfer load to regenerating tissue is an important part of the musculoskeletal healing process. This characteristic is only found in resorbable materials. Although metallic implants offer effective load-bearing properties in applications such as joint replacement and certain spinal surgeries, these high strength materials do not resorb and do not effectively transfer loads to the implant site. Due to the high strength of metals, these implants bear the majority of the force at the site and can shield the surrounding tissues from any load. This phenomenon is called stress shielding and can actually cause bone to resorb in certain areas around the implant (8,9). The stress-shielding effect is based on a concept called Wolff's Law, which describes the ability of bone to

dynamically respond to the presence or lack of stress by changing its density and strength.

When bone is subjected to new loads, the additional stress stimulates bone formation and the tissue increases in strength and density. When the remodeling process is complete, the stronger tissue can now fully support the added load. However, when a high strength material such as metal is placed in bone, the bone surrounding the implant is shielded from the normal stresses, which results in a decrease in the strength and density of this tissue and possible bone resorption. This phenomenon can cause complications such as implant loosening or fracture of the implant site. Polymer and ceramic materials, on the other hand, have mechanical properties that are similar to bone, which allows them to share the stresses with newly regenerating tissue thereby preventing resorption and other stress-shielding complications (10–12).

Although load transfer and strength retention are common properties of all resorbable implants, not all surgical sites heal at the same rate. In fracture fixation applications where bone-to-bone contact is maintained, healing can be as short as 6–8 weeks. However, in applications such as spinal fusion where significant amounts of tissue need to be formed in the intervertebral space, the healing process can take up to 6–12 months. Based on the dependence of implant function on the surgical site, the material choice becomes an important part of implant development. The challenge in designing an implant lies in choosing a material that correctly matches the function and strength requirements of the surgical application, which can be accomplished through a thorough understanding of the function of the implant, the load requirements of the implant site, and the properties of the material.

RESORBABLE POLYMERS

One of the most versatile materials used in orthopedic surgery are polymers. Polymers are a group of materials that are produced through a chemical reaction that results in a long chain of repeating molecules called monomers. In addition to polymers composed of a single monomer repeating unit, there are other materials, called copolymers, that have two or more monomer repeating units. By combining different monomers, the properties of the resultant copolymer can be specifically modified to serve a certain purpose. This versatility can also be achieved by modifying the polymerization reaction and the postprocessing techniques used to create polymer implants. Table 1 shows a few examples of the many properties that characterize polymers. These characteristics can be altered by changing the molecular weight, chemical structure, and morphology of the polymer or copolymer.

Table 1. Range of Common Properties Found in Orthopedic Polymers

Property	Range	
Resorbability	Fully Resorbable	Nonresorbable
Strength	Low Strength	High Strength
Moldability	Flexible	Rigid
Physical State	Gel/Liquid	Solid
Temperature Sensitivity	Flexible at higher Temperature	Rigid at all Temperatures
Radiation Resistance	Low	High

The molecular weight of a polymer is a measurement of the number of repeating units found in the entire molecule. During the formation of polymers and copolymers, the length of the molecule can be controlled to give a variety of molecular weights. The length of the polymer chain can be as small as a few thousand repeating units or as large as a million, which can have a significant effect on the degradation properties of the polymer. When a polymer breaks down, it occurs through random cleavage of the chemical bonds along the polymer chain. It is not until the polymer finally fragments into its monomer form that the material is absorbed by the surrounding tissue. Therefore, longer polymers chains with higher molecular weights will take a longer time to degrade because more bonds exist to the cleaved.

Additionally, the chemical structure can also affect degradation. As described previously, the backbone of a polymer consists of a long, continuous chain of monomer units linked together. In all resorbable polymers, it is the backbone of the polymer where degradation occurs. The typical linkage that allows polymers to break down is a carbon–oxygen–carbon (C-O-C) bond. This bond is found in ester, carbonate, carboxylic acid, and amide-based polymers. The degradation process occurs at this bond when the material is exposed to water. In a process called hydrolytic degradation, water molecules chemically react with the C-O-C bonds causing them to break apart at random areas throughout the polymer chain. The chemical structure of the polymer dictates the ability of the water molecules to access these bonds and start the degradation reaction. If the polymer is characterized by large bulky side chains or strong C-O-C bonds, it becomes difficult for the water molecule to penetrate the polymer chains to react with the backbone, which results in a prolonged degradation period. The opposite is true for polymers that tend to absorb water and do not have any large side chains. In these polymers, the water molecules can easily access the backbone and the degradation process proceeds at a relatively fast rate.

The final characteristic that can affect the degradation and strength of a polymer is the morphology. The morphology of the polymer refers to the orientation of the long polymer chains throughout the material. Polymer morphology can be classified into three groups: crystalline polymers, semicrystalline polymers, and amorphous polymers. The crystallinity of a polymer develops from areas within the material where the polymer chains are aligned and tightly packed together. This type of orientation forms dense crystalline regions within the random arrangement of the polymer chains. A highly organized polymer is considered crystalline, whereas a completely random orientation is considered amorphous. Semicrystalline polymers fall between these two extremes and exist with varying degrees of crystallinity (Fig. 2).

The effect of crystallinity on the degradation of the polymer is due to the tight orientation between the polymer chains in the crystalline regions. With highly crystalline

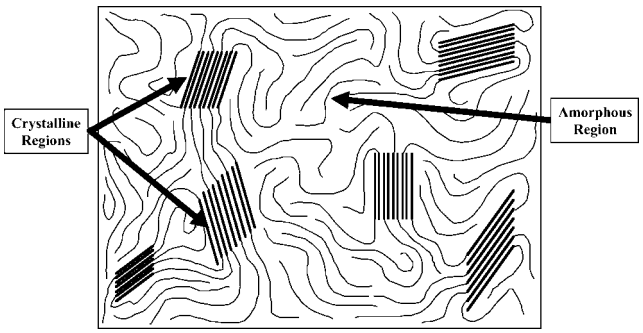


Figure 2. Semicrystalline polymer showing orientation of amorphous regions and crystalline regions.

polymers, the degradation rate is very slow due to the difficulty of water gaining access to the C-O-C bonds. These polymers degrade at a rate much slower than polymers that are completely amorphous with no crystalline regions (13). The crystallinity also affects the mechanical properties of the polymer. The dense, organized areas within crystalline polymer make these regions stronger than the unorganized, amorphous regions. As a result, an increase in crystallinity translates into an increase in mechanical properties.

The ability to alter the properties of a polymer has resulted in thousands of different materials used in a wide range of applications. However, only a few of these polymers can be effectively used as medical implants due to the strict requirements of surgical implants. The following sections describe some of the polymers currently used in orthopedic surgery.

Poly(hydroxy acids)

Poly(hydroxy acids) were the first group of resorbable materials to be used in surgery (14). The main polymers in this family are poly(lactic acid) (PLA), poly(glycolic acid) (PGA), and the copolymer poly(lactide-*co*-glycolide) (PLG). The basic chemical structure of these materials in shown in Fig. 3. Originally, PLA and PGA were initially used as a degradable sutures (15–18). However, since their initial success in the wound closure field, both of these polymers have been fabricated into several orthopedic implants including screws (19,20), plates (19,21), pins (22–25), suture anchors (26), and bone grafting scaffolds (27–30). In addition, several new devices composed of the PLG copolymer have been developed over the past 10 years (31–35).

Although the chemical structure of PLA and PGA is somewhat similar, the presence of a methyl group ($-CH_3$) in PLA significantly changes its physical properties compared with PGA. Comparatively, PGA has a lower strength and degrades in approximately 3–6 months, whereas certain forms of PLA can take 3–5 years to fully degrade. Although only a single methyl group differentiating PLA from PGA exists, the location of this side group close to the C-O-C bond makes it difficult for the water molecules to gain access to cleavage site, thereby prolonging degradation.

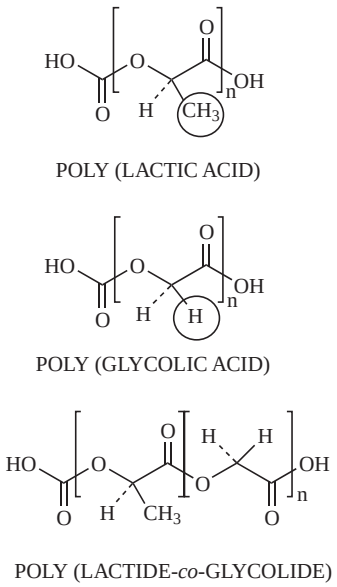


Figure 3. Chemical structure of poly(lactic acid), poly(glycolic acid), and the copolymer poly(lactide-*co*-glycolide).

In addition, the methyl group in PLA also gives the polymer a unique chemical orientation. As a monomer, lactic acid is a molecule that can have two different molecular orientations: L-lactic acid and D-lactic acid. These isomers are based on the orientation of the methyl and hydrogen groups on the molecule. Figure 4 shows the chiral nature of the lactic acid molecule and the resulting stereoregular polymers: poly(L-lactic acid) (PLLA), poly (D-lactic acid (PDLA), and poly (D,L-lactic acid) (PDLLA). Although three forms of PLA exist, in the medical field, poly(L-lactic acid) is used more often than poly(D-lactic acid) because the degradation product is the same as naturally occurring L-lactic acid (13).

Using the various forms of PLA, polymers with significantly different properties can be synthesized. The effect of the starting isomer on the physical properties of the material is dramatically seen in the properties of PLLA and PDLLA. In Fig. 4, the chemical structure of poly (L-lactic acid) is represented by a long chain with all of the $-CH_3$ groups on one side. This uniformity allows the chains to pack tightly together resulting in a highly crystalline material that has a high strength and long degradation period (3–5 years). Poly(D,L lactic acid), on the other hand, is characterized by either a random or alternating arrangement of the $-CH_3$ groups and –H groups. This molecule orientation prevents the polymer chains from packing together, resulting in a completely amorphous polymer with a lower strength and shorter degradation profile (9–12 months). In addition, the polymerization of L-lactic acid and D,L-lactic acid together results in a copolymer with properties in between PLLA and PDLLA. In recent years, the 70:30 combination of poly(L/D,L lactic acid) has gained popularity in orthopedic applications due to its ability to retain its strength for 9–12 months while being completely resorbed within 1.5–2 years (36–38). This copolymer appears to provide the best

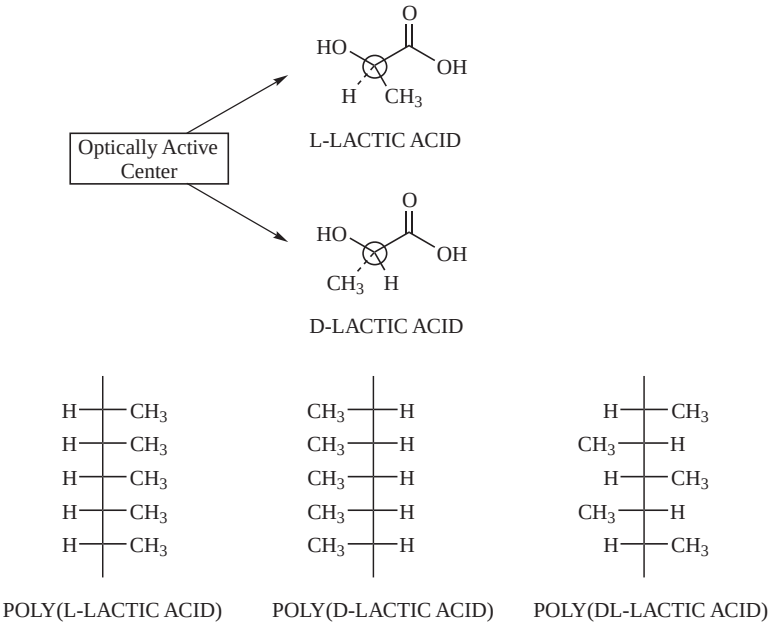


Figure 4. The optically active center in lactic acid is allows it to have two different molecular orientations. These orientations result in three types of stereoregular polymers.

of both worlds in that it has the strength retention of PLLA but has a degradation period only slightly longer than PDLLA.

In addition to the lactic acid-based copolymers, a combination of PLA and PGA has also been shown to be an effective implant material (31–35). Due to the large differences in the degradation properties of PLA and PGA, the poly(lactide-*co*-glycolide) (PLG) copolymer can be modified based on the PLA to PGA ratio to provide varying degradation periods. Common PLG copolymers used in orthopedic surgery have PLA/PGA ratios of 50:50, 75:25, and 85:15. This combination not only provides both slow and fast resorbing monomer units but also eliminates any crystallinity, making the copolymer completely amorphous. These materials have been commonly used as fracture implants due to the shorter 6–12 month degradation period.

Although PLA and PGA polymers have been successfully used in patients for several years, there have been certain cases where the abundance of acidic monomers at the site has caused inflammation and bone resorption (39–48). When PLA and PGA polymers near the end of degradation, they release lactic acid and glycolic acid, respectively. Although these degradation products can be metabolized by the body, if the surrounding tissue cannot absorb the acid in a timely manner, the build up of acid and resultant drop in pH at the implant site can cause bone to resorb. Historically, this effect has mainly been seen in the fast-resorbing PGA implants; however, a few cases have been reported with PLA (43,46,49,50). Although the bone resorption complication is detrimental to the healing of the implant site, the complication rate has been relatively low. In a review of over 2000 patients by Bostman, only 5% of the patients have shown implant-associated reactions (44).

Additionally, the copolymers PLG and PLDLLA have been shown to possess a more osteocompatible degradation profile due to a gradual release of the acidic byproducts (36,51–56), which has minimized acid dumping and the associated bone resorption complications. In a study by Eppley et al. (35), 1883 patients treated with PLG plates and screws for bone fixation in craniofacial procedures showed an implant-related complication rate of only 0.5%, which was well below the 5% rate reported by Bostman for PGA and PLA implants. Overall, the PLG and PLDLLA copolymers have been shown to be effective devices for fracture fixation, bone graft containment, and soft tissue fixation, and have begun to replace the outdated PLA and PGA devices (37,38,57,58).

Polycarbonates

Another group of resorbable polymers are the polycarbonates. Although the majority of the polymers and copolymers within the polycarbonate family are nonresorbable plastics used for industrial applications, a select few exist that are resorbable and can be used as orthopedic implants. One group of medical-grade polycarbonates are the copolymers based off of poly(trimethylene-carbonate) (PTMC) and poly(glycolic acid) or poly(lactic acid). These combinations offer the combined advantage of the processing versatility of PTMC and the resorbability and strength of PLA and PGA. The PTMC copolymers have been used for soft tissue fixation in shoulder surgery as suture anchors and soft tissue tacks (59–61).

Although the PTMC copolymers with PGA and PLA offer improved implant properties compared with PTMC alone, the degradation of the material still produces acidic monomers. In order to avoid the issues with glycolic acid- and lactic acid-based polymers and copolymers, an amino acid-based polycarbonate was developed by Joachim Kohn at Rutgers University. Designed specifically for orthopedic applications, the amino acid poly(carbonates) combine the biocompatibility of individual amino acids with the strength and processability of standard industrial poly(carbonates) (62–64). One such promising polymer, poly(DTE carbonate), is derived from the amino acid

tyrosine and has been shown to have excellent strength-retention properties, an optimal degradation profile, and biocompatible degradation products (65–68). Based on large amount of characterization data, a material safety file has been recently established at the U.S. Food and Drug Administration (FDA) that allows manufacturers to begin development of poly(DTE carbonate) implants. Due to the advantages of poly(DTE carbonate) over conventional resorbable polymers, amino acid-based poly(carbonate) implants may soon be a common sight in orthopedic operating rooms.

Other Resorbable Polymers

In addition to the widely used PLA and PGA polymers and the up-and-coming amino acid-based poly(carbonates), several other polymers have applications as medical devices. Although not specifically used in orthopedics, the poly(anhydride) family of polymers developed by Robert Langer at MIT has been effectively used as drug-delivery vehicles (69–73). The function of these resorbable implants is to provide a sustained and controlled release of drugs to a specific implant site. The device functions by releasing molecules entrapped within the implant as it degrades. Another polymer, poly(dioxanone), has been used as a resorbable suture material for several years (74–80). The flexibility of this polymer enables it to be used as a monofilament suture instead of the typical braided fiber of PGA, which provides the suture with an improved ability to move through tissue with less friction, thereby minimizing the tearing and pulling of the surrounding areas (81,82). Looking specifically at orthopedic applications, additional polymers currently in development include poly(caprolactone) (83–86), poly(hydroxybutyrate) (87–89), polyurethanes (90–93), and poly(phosphazenes) (94–96).

RESORBABLE CALCIUM CERAMICS

Aside from the polymers, the other group of resorbable implant materials are the calcium-based ceramics. Due to the similarity of these materials with the mineral content of bone, hydroxyapatite [$Ca_{10}(PO_4)_6(OH)_2$], calcium ceramics are highly biocompatible and osteocompatible materials that have a long history of clinical use. These materials are typically used in orthopedic surgery to fill voids in bone as self-setting cements or as porous blocks and granules.

Calcium Sulfate

One the first materials to ever be used as a filler for bone defects was calcium sulfate (Plaster of Paris) (97). In its dehydrated form (calcium sulfate hemihydrate), this material undergoes a chemical reaction when mixed with water that allows it to function as a resorbable cement. As the cement reacts, it transforms from a slurry, to a paste, to a dough, and then fully sets into its final hardened form (calcium sulfate dihydrate). This reaction is exothermic in that it produces heat; however, the increase in temperature is only slightly above body temp (37 °C). Figure 5 shows a typical timeline of the calcium sulfate setting reaction. In the slurry and paste form, the calcium sulfate is able to be added to a syringe and injected to the bone graft site. Near the end of the reaction, the cement becomes much thicker and has a putty-like consistency. During this phase, the doughy cement can be molded into a variety of shapes and provides a custom fit when placed directly at the implant site. Once the cement has fully hardened, it can be shaped by using powered surgical instruments such as osteotomes, burrs, and drills.

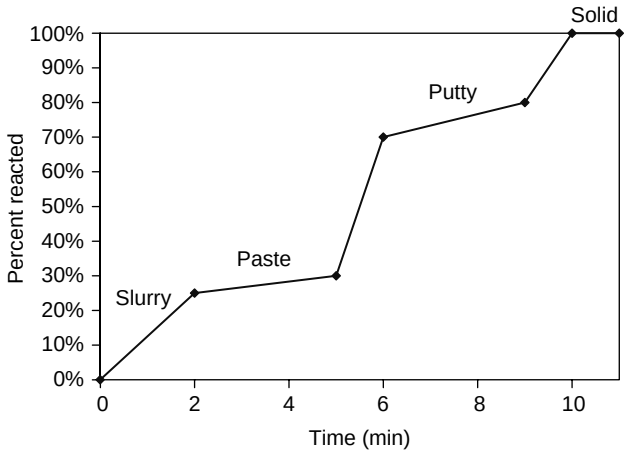


Figure 5. Typical setting reaction and phase changes for a calcium sulfate cement.

The resorption of calcium sulfate graft materials is based on the microstructure of the fully hardened cement. Figure 6 shows electron micrographs of the surface of fully reacted calcium sulfate dihydrate. These high magnification images show small calcium sulfate crystals packed together in a microporous structure. Upon implantation, the presence of these small pores allows the calcium sulfate to absorb water throughout the cement. Unlike polymers, which undergo active breakdown of the polymer chains, calcium sulfate materials are slowly dissolved by the water. As the material dissolves, Ca^{2+} and SO_4^{-3} ions are released over a 6–8 week period. During healing, bone formation initially begins on the outer area of the calcium sulfate and progresses inward as the cement slowly breaks apart. During the resorption process, the dissolution of the calcium sulfate material aids bone formation by providing a direct source Ca^{2+} ions to the surrounding osteoblasts. These cells absorb the calcium and use it during the mineralization phase of bone regeneration. From a mechanical standpoint, the hardened cement can provide initial stabilization to the site, but quickly loses it strength as the calcium sulfate begins to fragment. Although the strength of the calcium sulfate quickly decreases within the first few weeks, additional bone regeneration takes place within the cement and the implant site becomes mechanically stable. At the 6–8 week period, the majority of the calcium sulfate is resorbed by the body and has been replaced by bone.

In general, calcium sulfate cements and implants offer an effective means to fill small voids in bone resulting from cysts, tumors, or fractures (98–101). The initial strength

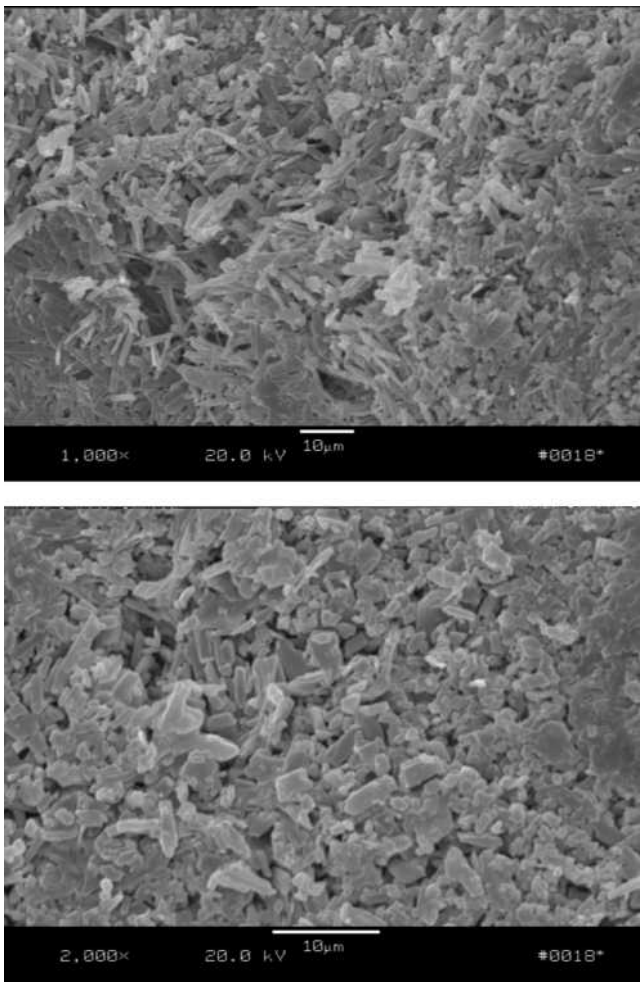


Figure 6. High magnification scanning electron micrographs of fully reacted calcium sulfate dihydrate showing crystalline structure and microporosity (1000X and 2000X magnification).

can also help maintain the spacing of fracture fragments and aid in placement of additional hardware. The moldability of the cement allows a custom fit to the defect site and makes the material easy to use. However, due to the quick resorption time and quick loss in strength, this material can not be effectively used in large defects or in areas under high mechanical loads. In these applications, supplemental hardware and grafting materials are needed to ensure complete bone regeneration (102,103). From a commercial standpoint, calcium sulfate graft materials are available in a cement form (requires mixing at the time of surgery) or in a preformed pellet form (fully reacted calcium sulfate dihydrate).

Calcium Phosphate. Calcium phosphates are another class of calcium containing bone graft materials that offer different properties than the calcium sulfates. As the name describes, these material are composed of varying amounts of calcium (Ca^{2+}) and phosphate (PO_4^{-3}). One of the first calcium phosphate materials to be used as a bone graft was hydroxyapatite, which was chosen because it is the main inorganic component of bone accounting for 40% of its weight. Most calcium phosphate graft materials are produced synthetically and can be chemically altered to create materials with different properties. By slightly varying the calcium-to-phosphate ratio, the resorption times and mechanical properties of these materials can be significantly altered. Hydroxyapatite [$Ca_{10}(PO_4)_6(OH)_2$] with a Ca/P ratio of 1.67 has slow resorption rate, which, depending on crystallinity, can be as little are 2–5% resorption per year. Tricalcium phosphate $Ca_3(PO_4)_2$ has a ratio of 1.5, which results in a much faster resorption time of 9–12 months.

Due to the chemical composition of calcium phosphates, the mechanism of resorption is different than the dissolution mechanism seen with calcium sulfates. The chemical similarity of calcium phosphates to bone results in a cell-mediated resorption profile. During healing, bone-resorbing cells called osteoclasts migrate to the surface of the calcium phosphate ceramics. Once activated, the osteoclasts release specific enzymes that dissolve the calcium phosphate into its base ions. As the osteoclasts tunnel through the calcium phosphate, bone-forming cells called osteoblasts trail behind filling in the region with new tissue. Similar to calcium sulfate, the calcium ions resulting from the resorption process are transported to the osteoblasts, which create new mineralized bone. Over time, the entire structure is slowly dissolved by the osteoclasts and replaced with new bone.

To facilitate this type of resorption process, many of the calcium phosphate bone graft materials exist as porous scaffolds (104–109). A typical example of an osteoconductive calcium phosphate bone graft scaffold is shown in Fig. 7. This material, called Pro Osteon (developed and manufactured by Interpore Cross), was one of the first porous calcium phosphates used in orthopedics (110–113). Derived from sea coral, it is fabricated by chemically converting the calcium carbonate skeleton of the coral into hydroxyapatite. This reaction can be run to completion to give a implant composed entirely of hydroxyapatite or intentionally stopped to result in an implant with a thin (4–10 μm) surface of hydroxyapatite over the calcium carbonate skeleton. The conversion of coral to Pro Osteon

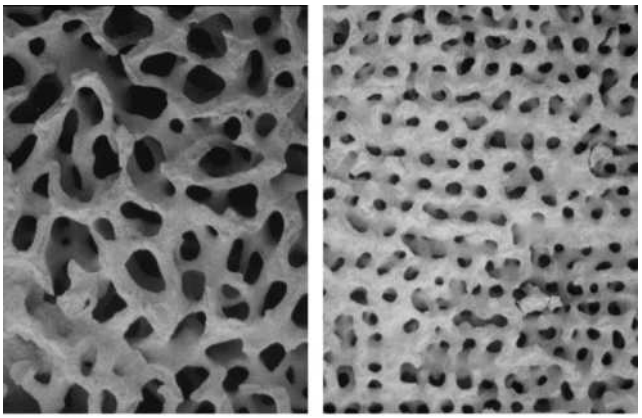


Figure 7. Photographs of a two commercially available calcium phosphate scaffolds derived from coral (Interpore Cross, Irvine, CA).

allows the relatively short degradation time of calcium carbonate (6–8 weeks) to be prolonged to 8–18 months for Pro Osteon R (HA layer on the calcium carbonate skeleton) and to 3–5 years for Pro Osteon HA (fully converted hydroxyapatite). With a natural pore structure similar to cancellous bone, the Pro Osteon graft materials offer an effective scaffold for new bone growth. Since development of the Pro Osteon bone graft materials, several other porous calcium phosphates have entered the market. These materials are synthetically made to mimic the porosity of cancellous bone, which is done through various foaming and void creation techniques.

In contrast to calcium sulfate graft materials, the slower resorption profile of porous calcium phosphate ceramics allow these material to be used in larger defects. In this scenario, the graft serves as a cellular "bridge" for continued bone growth. In bone grafting surgery, once a defect reaches a size when it can no longer completely heal itself, it is called a critical-sized defect. Typical bone regeneration can bridge empty gaps of up to 4 mm, but anything larger will not fill in with bone. A porous ceramic scaffold alleviates this problem by providing the means for bone to grow across the entire defect.

This effect was demonstrated in a study by Holmes who implanted a block of Pro Osteon 500R (calcium carbonate scaffold with an HA coating) into a rabbit tibial defect (114). The healing sequence of the this scaffold is shown in Fig. 8. As seen from cross sectional image of the implant before implantation (Fig. 8a), the structure is characterized by an open pore structure (black regions) within areas of calcium carbonate/HA ceramic (light-gray regions). After initial placement of the porous ceramic, cells migrated to the graft site and began to infiltrate the pore system. At the same time, proteins were released from surrounding bone and blood cells to stimulate the bone regeneration process, which was seen in the 6 week histology of the Pro Osteon 500R implant (Fig. 8b). In this image, bone formation was evident within the porosity of the scaffold, and osteoclasts were seen resorbing the scaffold (arrows). By 12 weeks, further bone growth was seen within the porosity, and significant portions of the scaffold were replaced by bone. At the 24 week time point, the scaffold was fully replaced by bone with the exception of the thin HA layer that once covered the calcium carbonate. As seen from this study, porous ceramics are capable of functioning as a scaffold for bone growth. The pore system allowed for immediate bone regeneration and the resorbability allowed the implant to be completely replaced by bone.

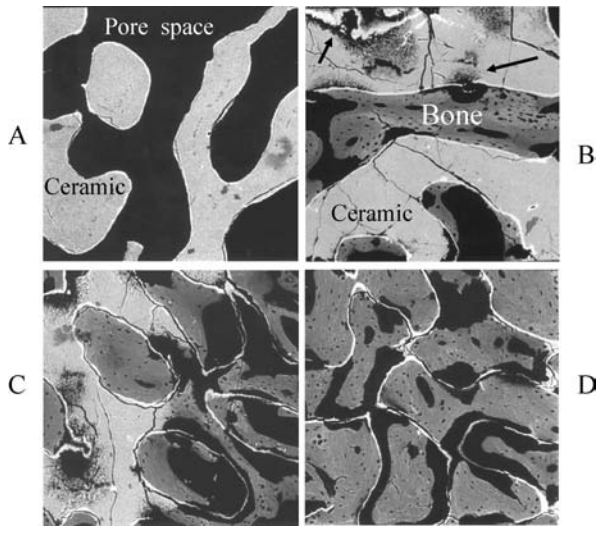


Figure 8. Typical healing mechanism of a porous ceramic implanted into a rabbit tibial defect. (Figure A – 0 weeks; Figure B – 6 weeks; Figure C – 12 weeks; Figure D – 24 weeks).

In addition to porous blocks and granules, calcium phosphates are also used in cement form (115–119). In this application, the base components that create calcium phosphates are provided in an unreacted form. With the addition of water, dilute acid, or other initiators, a chemical reaction takes place, and the components are converted to calcium phosphate. The result is a moldable paste or putty that can be shaped to the graft site and hardens into a solid mass. Although these cements have longer resorption times than calcium sulfate cements and can be used in broader applications, the resulting hardened cement does not possess the porosity to function as a scaffold for bone repair, which has limited the use of calcium phosphate cerments because surgeons prefer the porous blocks and granules over the self-setting cements.

RESORBABLE COMPOSITES

As discussed, both polymers and ceramics have properties suitable for fabricating orthopedic implants. However, certain drawbacks exist with these materials that can not be avoided no matter how the material is fabricated or chemically altered. One technique for combining the desirable properties of two or more materials is the fabrication of a composite. Composites used in the medical device area are fabricated by physically mixing two or more resorbable materials. One of the most common composite combinations is the creation of a polymer-ceramic composite. On their own, ceramics are excellent substrates for new bone growth due to the chemical similarity with bone mineral. However, their brittleness limits their use in load-bearing applications. Polymers, on the other hand, are elastomeric materials that can flex under deformation without major structural collapse. The combination of these two materials results in a high strength, yet ductile composite that allows for direct bone attachment on its surface. In this combination, the polymer adds to the overall mechanical properties of the composite, whereas the ceramic allows for bone formation directly on the ceramic phase.

The fabrication of a composite is a relatively straightforward process. Typically, ceramic particles in the shape of spheres or fibers are added to the polymer during processing. The various orientations of the particles within a polymer are shown in Fig. 9. As seen from these illustrations, each particle is surrounded by the polymer and serves to reinforce the polymer phase and improve it mechanical properties. Once fabricated in a block or rod

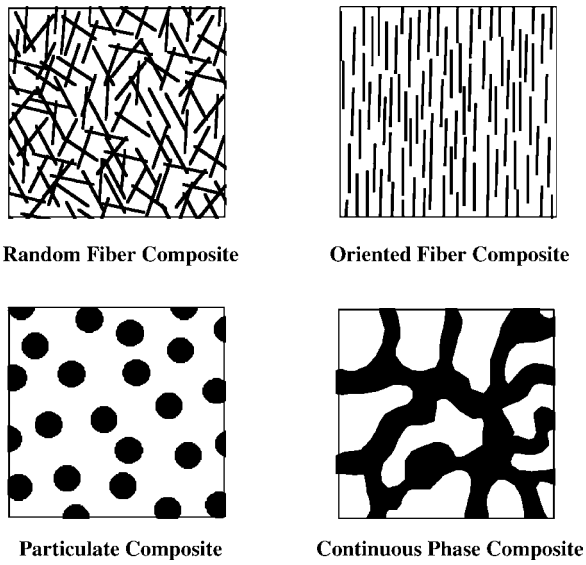


Figure 9. Orientation of various types of polymer-ceramic composites (ceramic is depicted as the black particles).

form, composites of these different types can be machined into a variety of implants such as fracture screws, pins, and plates. During the machining process, the ceramic on the outer surfaces of the implant are exposed. From a bone implant standpoint, the presence of the exposed calcium ceramic particles on the surface of the polymer aids in creating a solid bone-to-implant interface. In comparison, pure polymer implants typically heal with limited bone contact or a continuous layer of fibrous tissue usually covering the surface. Although the implant can still provide stabilization, it is not directly bonded to the surrounding bone. A composite implant improves on the stabilizing effect of the device through this bone-bonding ability.

In addition to the particulate ceramic composites, a new type of composite has recently been developed by Interpore Cross (Irvine, CA). This novel material consists of two intact, continuous phases of polymer and ceramic. Shown in Fig. 10, a continuous phase composite (CPC) is the result of infiltrating a porous ceramic block with polymer. The

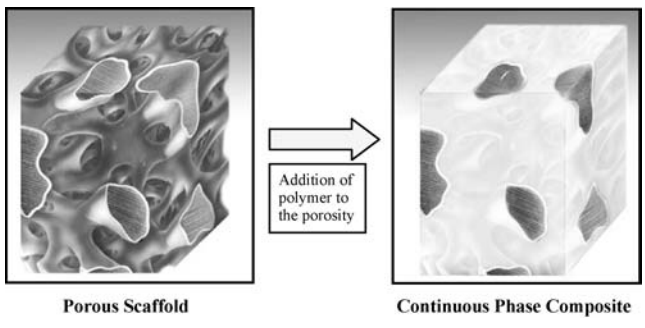


Figure 10. A continuous phase composite is formed when a polymer is infiltrated into the porosity a porous, ceramic scaffold. The result is a solid block with an intact polymer and ceramic phase.

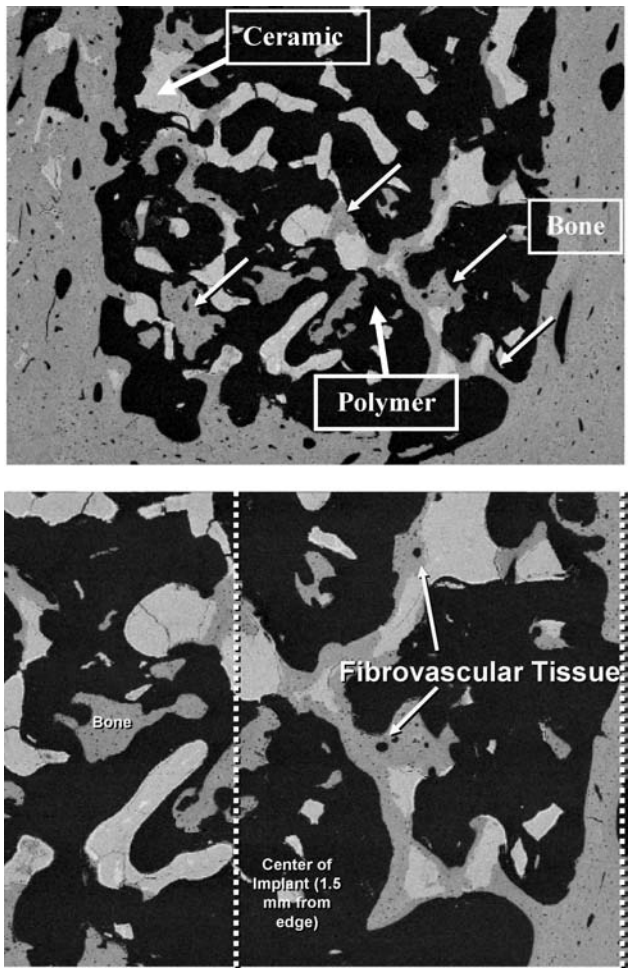


Figure 11. Backscattered electron microscope images demonstrating bone and blood vessel in-growth into a CPC implant (bone is shown in gray, ceramic is white, and polymer is black).

end result is a composite material with continuous seams of ceramic running through the polymer. Similar to the particulate composites, the CPC material will allow for bone growth on the surface and into the ceramic regions. However, the continuity of the ceramic phase throughout the composite gives the material a unique ability to allow for bone to penetrate into the center of a CPC implant. Figure 11 shows the histology a CPC implant composed of the Pro Osteon porous ceramic infiltrated with poly(L/D,L lactic acid) implanted in a sheep femur at 9 months. This backscattered electron microscope image shows the ability of a CPC implant to support bone and blood vessel in-growth into the center of the implant wall. In addition to acting as a structural implant, a CPC device also functions as an eventual scaffold for bone in-growth. From a healing standpoint, this type of composite will result in more bone formation at the site. Additionally, the presence of bone and blood vessels within the implant wall significantly improves the ability of the tissue to absorb the degradation products. Typically occurring at the surface of polymer implants, the presence of bone within the CPC material allows resorption throughout the entire device. This new

composite is currently being investigated for use as spinal fusion implants, fracture screws and plates, interference screws, and suture anchors.

CONCLUSION

As seen from this article, resorbable polymers and ceramics possess the desired properties needed for orthopedic implants. They have been shown to be versatile materials with a range in degradation rates and mechanical properties. The resorbable nature of these devices allows them to provide temporary stabilization and mechanical support. Combined with the ability to be completely resorbed by the body and replaced by natural tissue, these implants are highly desirable alternatives to their nonresorbing counterparts. The elimination of long-term implant complications and the ability to share load with regenerating tissues are large driving forces behind the use of these implants in orthopedics. With further advancements in biomaterial research, resorbable implants may soon become the standard of care.

BIBLIOGRAPHY

Cited References

1. History of Total Joint Replacement. Utah Hip and Knee Center. Available: http://www.utahhipandknee.com/history.htm. Accessed September 24, 2004.
2. History of Innovation. Zimmer Corporate Website. Available:http://www.zimmer.com/ctl?op=global&action=1&id=1420&template= CP. Accessed September 24, 2004.
3. Parker DA, Dunbar MJ, Rorabeck CH. Extensor mechanism failure associated with total knee arthroplasty: Prevention and management. J Am Acad Orthop Surg 2003;11:238–247.
4. Ries MD. Complications in primary total hip arthroplasty: Avoidance and management: wear. Instrum Course Lect 2003;52:257–265.
5. Sochart DH. Relationship of acetabular wear to osteolysis and loosening in total hip arthroplasty. Clin Orthop 1999; 135–150.
6. Moreland JR. Mechanisms of failure in total knee arthroplasty. Clin Orthop 1988; 49–64.
7. Borden MD. The development of bone graft materials using various formulations of demineralized bone matrix. In Laurencin CT, editor, Bone Graft Substitutes. Conshohocken, (PA): ASTM International; 2003.
8. Uhthoff HK, Finnegan M. The effects of metal plates on post-traumatic remodelling and bone mass. J Bone Joint Surg Br 1983;65:66–71.
9. van der EM, Dijkema AR, Klein CP, Patka P, Haarman HJ. Tissue reaction on PLLA versus stainless steel interlocking nails for fracture fixation: An animal study. Biomaterials 1995;16:103–106.
10. Bos RR, Rozema FR, Boering G, Nijenhuis AJ, Pennings AJ, Verwey AB. Bio-absorbable plates and screws for internal fixation of mandibular fractures. A study in six dogs. Int J Oral Maxillofac Surg 1989;18:365–369.
11. Viljanen J, Kinnunen J, Bondestam S, Majola A, Rokkanen P, Tormala P. Bone changes after experimental osteotomies fixed with absorbable self-reinforced poly-L-lactide screws or metallic screws studied by plain radiographs, quantitative computed tomography and magnetic resonance imaging. Biomaterials 1995;16:1353–1358.
12. van der EM, Klein CP, Blieck-Hogervorst JM, Patka P, Haarman HJ. Bone tissue response to biodegradable polymers used for intra medullary fracture fixation: A long-term in vivo study in sheep femora. Biomaterials 1999; 20:121–128.
13. Engelberg I, Kohn J. Physico-mechanical properties of degradable polymers used in medical applications: A comparative study. Biomaterials 1991;12:292–304.
14. Kulkarni RK, Pani KC, Neuman C, Leonard F. Polylactic acid for surgical implants. Arch Surg 1966;93:839–843.
15. Chu CC. A comparison of the effect of pH on the biodegradation of two synthetic absorbable sutures. Ann Surg 1982;195:55–59.
16. Chu CC. Mechanical properties of suture materials: An important characterization. Ann Surg 1981;193:365–371.
17. Chu CC, Campbell ND. Scanning electron microscopic study of the hydrolytic degradation of poly(glycolic acid) suture. J Biomed Mater Res 1982;16:417–430.
18. Chu CC. The in-vitro degradation of poly(glycolic acid) sutures–effect of pH. J Biomed Mater Res 1981;15:795–804.
19. Leenslag JW, Pennings AJ, Bos RR, Rozema FR, Boering G. Resorbable materials of poly(L-lactide). VI. Plates and screws for internal fracture fixation. Biomaterials 1987;8:70–73.
20. Tuompo P, Partio EK, Jukkala-Partio K, Pohjonen T, Helevirta P, Rokkanen P. Comparison of polylactide screw and expansion bolt in bioabsorbable fixation with patellar tendon bone graft for anterior cruciate ligament rupture of the knee. A preliminary study. Knee Surg Sports Traumatol Arthrosc 1999;7:296–302.
21. Koskikare K, Hirvensalo E, Patiala H, Rokkanen P, Pohjonen T, Tormala P, Lob G. Fixation of osteotomies of the distal femur with absorbable, self-reinforced, poly-L-lactide plates. An experimental study on rabbits. Arch Orthop Trauma Surg 1997;116:352–356.
22. Makela EA, Vainionpaa S, Vihtonen K, Mero M, Laiho J, Tormala P, Rokkanen P. Healing of physeal fracture after fixation with biodegradable self-reinforced polyglycolic acid pins. An experimental study on growing rabbits. Clin Mater 1990;5:1–12.
23. Pihlajamaki H, Bostman O, Hirvensalo E, Tormala P, Rokkanen P. Absorbable pins of self-reinforced poly-L-lactic acid for fixation of fractures and osteotomies. J Bone Joint Surg Br 1992;74:853–857.
24. Tormala P, Vasenius J, Vainionpaa S, Laiho J, Pohjonen T, Rokkanen P. Ultra-high-strength absorbable self-reinforced polyglycolide (SR-PGA) composite rods for internal fixation of bone fractures: In vitro and in vivo study. J Biomed Mater Res 1991;25:1–22.
25. Vainionpaa S, Kilpikari J, Laiho J, Helevirta P, Rokkanen P, Tormala P. Strength and strength retention in vitro, of absorbable, self-reinforced polyglycolide (PGA) rods for fracture fixation. Biomaterials 1987;8:46–48.
26. Barber FA, Deck MA. The in vivo histology of an absorbable suture anchor: A preliminary report. Arthroscopy 1995;11:77–81.
27. Borden M, El Amin SF, Attawia M, Laurencin CT. Structural and human cellular assessment of a novel microsphere-based tissue engineered scaffold for bone repair. Biomaterials 2003;24:597–609.
28. Borden M, Attawia M, Laurencin CT. The sintered microsphere matrix for bone tissue engineering: in vitro osteoconductivity studies. J Biomed Mater Res 2002;61:421–429.
29. Borden M, Attawia M, Khan Y, Laurencin CT. Tissue engineered microsphere-based matrices for bone repair: design and evaluation. Biomaterials 2002;23:551–559.

30. Thomson RC, Yaszemski MJ, Powers JM, Mikos AG. Fabrication of biodegradable polymer scaffolds to engineer trabecular bone. J Biomater Sci Polym Ed 1995;7:23–38.
31. Hollier LH, Rogers N, Berzin E, Stal S. Resorbable mesh in the treatment of orbital floor fractures. J Craniofac Surg 2001;12:242–246.
32. Edwards RC, Kiely KD, Eppley BL. The fate of resorbable poly-L-lactic/polyglycolic acid (LactoSorb) bone fixation devices in orthognathic surgery. J Oral Maxillofac Surg 2001;59:19–25.
33. Edwards RC, Kiely KD, Eppley BL. Resorbable PLLA-PGA screw fixation of mandibular sagittal split osteotomies. J Craniofac Surg 1999;10:230–236.
34. Westermark A. LactoSorb resorbable osteosynthesis after sagittal split osteotomy of the mandible: A 2-year follow-up. J Craniofac Surg 1999;10:519–522.
35. Eppley BL, Morales L, Wood R, Pensler J, Goldstein J, Havlik RJ, Habal M, Losken A, Williams JK, Burstein F, Rozzelle AA, Sadove AM. Resorbable PLLA-PGA plate and screw fixation in pediatric craniofacial surgery: Clinical experience in 1883 patients. Plast Reconstr Surg 2004;114:850–856.
36. Toth JM, Wang M, Scifert JL, Cornwall GB, Estes BT, Seim HB, III, Turner AS. Evaluation of 70/30 D,L-PLa for use as a resorbable interbody fusion cage. Orthopedics 2002;25: s1131–s1140.
37. Vaccaro AR, Madigan L. Spinal applications of bioabsorbable implants. Orthopedics 2002;25:s1115–s1120.
38. Lowe TG, Coe JD. Bioresorbable polymer implants in the unilateral transforaminal lumbar interbody fusion procedure. Orthopedics 2002;25:s1179–s1183.
39. Bostman OM. Osteolytic changes accompanying degradation of absorbable fracture fixation implants. J Bone Joint Surg Br 1991;73:679–682.
40. Pelto-Vasenius K, Hirvensalo E, Vasenius J, Rokkanen P. Osteolytic changes after polyglycolide pin fixation in chevron osteotomy. Foot Ankle Int 1997;18:21–25.
41. Bostman OM. Reaction to biodegradable implants. J Bone Joint Surg Br 1993;75:336–337.
42. Bostman OM. Intense granulomatous inflammatory lesions associated with absorbable internal fixation devices made of polyglycolide in ankle fractures. Clin Orthop 1992; 193–199.
43. Bostman OM. Osteoarthritis of the ankle after foreign-body reaction to absorbable pins and screws: A three- to nine-year follow-up study. J Bone Joint Surg Br 1998;80:333–338.
44. Bostman OM, Pihlajamaki HK. Adverse tissue reactions to bioabsorbable fixation devices. Clin Orthop 2000; 216–227.
45. Bostman O, Pihlajamaki H. Clinical biocompatibility of biodegradable orthopedic implants for internal fixation: A review. Biomaterials 2000;21:2615–2621.
46. Rovinsky D, Nissen TP, Otsuka NY. Osteolytic reaction to polylevolactic acid fracture fixation. Orthopedics 2001;24: 177–179.
47. Bostman O, Hirvensalo E, Makinen J, Rokkanen P. Foreign-body reactions to fracture fixation implants of biodegradable synthetic polymers. J Bone Joint Surg Br 1990;72:592–596.
48. Taylor MS, Daniels AU, Andriano KP, Heller J. Six bioabsorbable polymers: in vitro acute toxicity of accumulated degradation products. J Appl Biomater 1994;5:151–157.
49. Bergsma JE, de Bruijn WC, Rozema FR, Bos RR, Boering G. Late degradation tissue response to poly(L-lactide) bone plates and screws. Biomaterials 1995;16:25–31.
50. Bergsma EJ, Rozema FR, Bos RR, de Bruijn WC. Foreign body reactions to resorbable poly(L-lactide) bone plates and screws used for the fixation of unstable zygomatic fractures. J Oral Maxillofac Surg 1993;51:666–670.
51. Lanman TH, Hopkins TJ. Lumbar interbody fusion after treatment with recombinant human bone morphogenetic protein-2 added to poly(L-lactide-*co*-D,L-lactide) bioresorbable implants. Neurosurg Focus 2004;16:E9.
52. Couture DE, Branch CL, Jr. Posterior lumbar interbody fusion with bioabsorbable spacers and local autograft in a series of 27 patients. Neurosurg Focus 2004;16:E8.
53. Vaccaro AR, Robbins MM, Madigan L, Albert TJ, Smith W, Hilibrand AS. Early findings in a pilot study of anterior cervical fusion in which bioabsorbable interbody spacers were used in the treatment of cervical degenerative disease. Neurosurg Focus 2004;16:E7.
54. Cornwall GB, Ames CP, Crawford NR, Chamberlain RH, Rubino AM, Seim HB, III, Turner AS. In vivo evaluation of bioresorbable polylactide implants for cervical graft containment in an ovine spinal fusion model. Neurosurg Focus 2004;16:E5.
55. Lippman CR, Hajjar M, Abshire B, Martin G, Engelman RW, Cahill DW. Cervical spine fusion with bioabsorbable cages. Neurosurg Focus 2004;16:E4.
56. Robbins MM, Vaccaro AR, Madigan L. The use of bioabsorbable implants in spine surgery. Neurosurg Focus 2004;16:E1.
57. Ames CP, Crawford NR, Chamberlain RH, Cornwall GB, Nottmeier E, Sonntag VK. Feasibility of a resorbable anterior cervical graft containment plate. Orthopedics 2002;25: s1149–s1155.
58. DiAngelo DJ, Kitchel S, McVay BJ, Scifert JL, Cornwall GB. Bioabsorbable anterior lumbar plate fixation in conjunction with anterior interbody fusion cages. Orthopedics 2002;25: s1157–s1165.
59. Speer KP, Warren RF, Pagnani M, Warner JJ. An arthroscopic technique for anterior stabilization of the shoulder with a bioabsorbable tack. J Bone Joint Surg Am 1996;78:1801–1807.
60. Warner JJ, Miller MD, Marks P, Fu FH. Arthroscopic Bankart repair with the Suretac device. Part I: Clinical observations. Arthroscopy 1995;11:2–13.
61. Warner JJ, Miller MD, Marks P. Arthroscopic Bankart repair with the Suretac device. Part II: Experimental observations. Arthroscopy 1995;11:14–20.
62. Bourke SL, Kohn J. Polymers derived from the amino acid L-tyrosine: polycarbonates, polyarylates and copolymers with poly(ethylene glycol). Adv Drug Deliv Rev 2003;55:447–466.
63. Ertel SI, Kohn J. Evaluation of a series of tyrosine-derived polycarbonates as degradable biomaterials. J Biomed Mater Res 1994;28:919–930.
64. Ertel SI, Kohn J, Zimmerman MC, Parsons JR. Evaluation of poly(DTH carbonate), a tyrosine-derived degradable polymer, for orthopedic applications. J Biomed Mater Res 1995;29:1337–1348.
65. Choueka J, Charvet JL, Koval KJ, Alexander H, James KS, Hooper KA, Kohn J. Canine bone response to tyrosine-derived polycarbonates and poly(L-lactic acid). J Biomed Mater Res 1996;31:35–41.
66. Tangpasuthadol V, Pendharkar SM, Peterson RC, Kohn J. Hydrolytic degradation of tyrosine-derived polycarbonates, a class of new biomaterials. Part II: 3-yr study of polymeric devices. Biomaterials 2000;21:2379–2387.
67. Chaikof EL, Matthew H, Kohn J, Mikos AG, Prestwich GD, Yip CM. Biomaterials and scaffolds in reparative medicine. Ann N Y Acad Sci 2002;961:96–105.
68. Kohn J, Langer R. Poly(iminocarbonates) as potential biomaterials. Biomaterials 1986;7:176–182.

69. Ibim SM, Uhrich KE, Bronson R, El Amin SF, Langer RS, Laurencin CT. Poly(anhydride-*co*-imides): in vivo biocompatibility in a rat model. Biomaterials 1998;19:941–951.
70. Ibim SE, Uhrich KE, Attawia M, Shastri VR, El Amin SF, Bronson R, Langer R, Laurencin CT. Preliminary in vivo report on the osteocompatibility of poly(anhydride-co-imides) evaluated in a tibial model. J Biomed Mater Res 1998;43:374–379.
71. Uhrich KE, Ibim SE, Larrier DR, Langer R, Laurencin CT. Chemical changes during in vivo degradation of poly(anhydride-imide) matrices. Biomaterials 1998;19: 2045–2050.
72. Katti DS, Lakshmi S, Langer R, Laurencin CT. Toxicity, biodegradation and elimination of polyanhydrides. Adv Drug Deliv Rev 2002;54:933–961.
73. Attawia MA, Uhrich KE, Botchwey E, Langer R, Laurencin CT. In vitro bone biocompatibility of poly (anhydride-co-imides) containing pyromellitylimidoalanine. J Orthop Res 1996;14:445–454.
74. Ray JA, Doddi N, Regula D, Williams JA, Melveger A. Polydioxanone (PDS), a novel monofilament synthetic absorbable suture. Surg Gynecol Obstet 1981;153:497–507.
75. Ping OC, Cameron RE. The hydrolytic degradation of polydioxanone (PDSII) sutures. Part I: Morphological aspects. J Biomed Mater Res 2002;63:280–290.
76. Ping OC, Cameron RE. The hydrolytic degradation of polydioxanone (PDSII) sutures. Part II: Micromechanisms of deformation. J Biomed Mater Res 2002;63:291–298.
77. Ray JA, Doddi N, Regula D, Williams JA, Melveger A. Polydioxanone (PDS), a novel monofilament synthetic absorbable suture. Surg Gynecol Obstet 1981;153:497–507.
78. Bartholomew RS. PDS (polydioxanone suture): A new synthetic absorbable suture in cataract surgery. A preliminary study. Ophthalmologica 1981;183:81–85.
79. Lerwick E. Studies on the efficacy and safety of polydioxanone monofilament absorbable suture. Surg Gynecol Obstet 1983;156:51–55.
80. Cohen EL, Kirschenbaum A, Glenn JF. Preclinical evaluation of PDS (polydioxanone) synthetic absorbable suture vs chromic surgical gut in urologic surgery. Urology 1987;30:369–372.
81. Apt L, Henrick A. "Tissue-drag" with polyglycolic acid (Dexon) and polyglactin 910 (Vicryl) sutures in strabismus surgery. J Pediatr Ophthalmol 1976;13:360–364.
82. Homsy CA, McDonald KE, Akers WW, Short C, Freeman BS. Surgical suture-canine tissue interaction for six common suture types. J Biomed Mater Res 1968;2:215–230.
83. Rhee SH, Lee YK, Lim BS, Yoo JJ, Kim HJ. Evaluation of a novel poly(epsilon-caprolactone)-organosiloxane hybrid material for the potential application as a bioactive and degradable bone substitute. Biomacromolecules 2004;5: 1575–1579.
84. Rhee SH. Bone-like apatite-forming ability and mechanical properties of poly(epsilon-caprolactone)/silica hybrid as a function of poly(epsilon-caprolactone) content. Biomaterials 2004;25:1167–1175.
85. Ciapetti G, Ambrosio L, Savarino L, Granchi D, Cenni E, Baldini N, Pagani S, Guizzardi S, Causa F, Giunti A. Osteoblast growth and function in porous poly epsilon-caprolactone matrices for bone repair: a preliminary study. Biomaterials 2003;24:3815–3824.
86. Im SY, Cho SH, Hwang JH, Lee SJ. Growth factor releasing porous poly (epsilon-caprolactone)-chitosan matrices for enhanced bone regenerative therapy. Arch Pharm Res 2003;26:76–82.
87. Wang YW, Wu Q, Chen J, Chen GQ. Evaluation of three-dimensional scaffolds made of blends of hydroxyapatite and poly(3-hydroxybutyrate-*co*-3-hydroxyhexanoate) for bone reconstruction. Biomaterials 2005;26:899–904.
88. Yang M, Zhu S, Chen Y, Chang Z, Chen G, Gong Y, Zhao N, Zhang X. Studies on bone marrow stromal cells affinity of poly (3-hydroxybutyrate-*co*-3-hydroxyhexanoate). Biomaterials 2004;25:1365–1373.
89. Kose GT, Kenar H, Hasirci N, Hasirci V. Macroporous poly(3-hydroxybutyrate-*co*-3-hydroxyvalerate) matrices for bone tissue engineering. Biomaterials 2003;24:1949–1958.
90. Farso NF, Karring T, Gogolewski S. Biodegradable guide for bone regeneration. Polyurethane membranes tested in rabbit radius defects. Acta Orthop Scand 1992;63:66–69.
91. Gorna K, Gogolewski S. Preparation, degradation, and calcification of biodegradable polyurethane foams for bone graft substitutes. J Biomed Mater Res 2003;67A:813–827.
92. Grad S, Kupcsik L, Gorna K, Gogolewski S, Alini M. The use of biodegradable polyurethane scaffolds for cartilage tissue engineering: potential and limitations. Biomaterials 2003; 24:5163–5171.
93. Warrer K, Karring T, Nyman S, Gogolewski S. Guided tissue regeneration using biodegradable membranes of polylactic acid or polyurethane. J Clin Periodontol 1992;19:633–640.
94. Ambrosio AM, Sahota JS, Runge C, Kurtz SM, Lakshmi S, Allcock HR, Laurencin CT. Novel polyphosphazene-hydroxyapatite composites as biomaterials. IEEE Eng Med Biol Mag 2003;22:18–26.
95. Laurencin CT, El Amin SF, Ibim SE, Willoughby DA, Attawia M, Allcock HR, Ambrosio AA. A highly porous 3-dimensional polyphosphazene polymer matrix for skeletal tissue regeneration. J Biomed Mater Res 1996;30:133–138.
96. Laurencin CT, Norman ME, Elgendy HM, El Amin SF, Allcock HR, Pucher SR, Ambrosio AA. Use of polyphosphazenes for skeletal tissue regeneration. J Biomed Mater Res 1993;27:963–973.
97. Coetzee AS. Regeneration of bone in the presence of calcium sulfate. Arch Otolaryngol 1980;106:405–409.
98. Gitelis S, Piasecki P, Turner T, Haggard W, Charters J, Urban R. Use of a calcium sulfate-based bone graft substitute for benign bone lesions. Orthopedics 2001;24:162–166.
99. Ladd AL, Pliam NB. Use of bone-graft substitutes in distal radius fractures. J Am Acad Orthop Surg 1999;7:279–290.
100. Guarnieri R, Pecora G, Fini M, Aldini NN, Giardino R, Orsini G, Piattelli A. Medical grade calcium sulfate hemihydrate in healing of human extraction sockets: Clinical and histological observations at 3 months. J Periodontol 2004;75:902–908.
101. Kelly CM, Wilkins RM. Treatment of benign bone lesions with an injectable calcium sulfate-based bone graft substitute. Orthopedics 2004;27:s131–s135.
102. Urban RM, Turner TM, Hall DJ, Infanger S, Cheema N, Lim TH. Healing of large defects treated with calcium sulfate pellets containing demineralized bone matrix particles. Orthopedics 2003;26:s581–s585.
103. Turner TM, Urban RM, Hall DJ, Infanger S, Gitelis S, Petersen DW, Haggard WO. Osseous healing using injectable calcium sulfate-based putty for the delivery of demineralized bone matrix and cancellous bone chips. Orthopedics 2003;26:s571–s575.
104. Thalgott JS, Giuffre JM, Fritts K, Timlin M, Klezl Z. Instrumented posterolateral lumbar fusion using coralline hydroxyapatite with or without demineralized bone matrix, as an adjunct to autologous bone. Spine J 2001;1:131–137.
105. McConnell JR, Freeman BJ, Debnath UK, Grevitt MP, Prince HG, Webb JK. A prospective randomized comparison of coralline hydroxyapatite with autograft in cervical interbody fusion. Spine 2003;28:317–323.
106. Thalgott JS, Klezl Z, Timlin M, Giuffre JM. Anterior lumbar interbody fusion with processed sea coral (coralline

hydroxyapatite) as part of a circumferential fusion. Spine 2002;27:E518–E525.
107. Delecrin J, Takahashi S, Gouin F, Passuti N. A synthetic porous ceramic as a bone graft substitute in the surgical management of scoliosis: A prospective, randomized study. Spine 2000;25:563–569.
108. McAndrew MP, Gorman PW, Lange TA. Tricalcium phosphate as a bone graft substitute in trauma: Preliminary report. J Orthop Trauma 1988;2:333–339.
109. Bucholz RW, Carlton A, Holmes RE. Hydroxyapatite and tricalcium phosphate bone graft substitutes. Orthop Clin North Am 1987;18:323–334.
110. Holmes R, Mooney V, Bucholz R, Tencer A. A coralline hydroxyapatite bone graft substitute. Preliminary report. Clin Orthop 1984; 252–262.
111. Finn RA, Bell WH, Brammer JA. Interpositional "grafting" with autogenous bone and coralline hydroxyapatite. J Maxillofac Surg 1980;8:217–227.
112. Holmes RE. Bone regeneration within a coralline hydroxyapatite implant. Plast Reconstr Surg 1979;63:626–633.
113. Holmes RE, Salyer KE. Bone regeneration in a coralline hydroxyapatite implant. Surg Forum 1978;29:611–612.
114. Jamali A, Hilpert A, Debes J, Afshar P, Rahban S, Holmes R. Hydroxyapatite/calcium carbonate (HA/CC) vs. plaster of Paris: A histomorphometric and radiographic study in a rabbit tibial defect model. Calcif Tissue Int 2002;71:172–178.
115. Kenny SM, Buggy M. Bone cements and fillers: A review. J Mater Sci Mater Med 2003;14:923–938.
116. Horstmann WG, Verheyen CC, Leemans R. An injectable calcium phosphate cement as a bone-graft substitute in the treatment of displaced lateral tibial plateau fractures. Injury 2003;34:141–144.
117. Kamano M, Honda Y, Kazuki K, Yasudab M. Palmar plating with calcium phosphate bone cement for unstable Colles' fractures. Clin Orthop 2003; 285–290.
118. Zimmermann R, Gabl M, Lutz M, Angermann P, Gschwentner M, Pechlaner S. Injectable calcium phosphate bone cement Norian SRS for the treatment of intra-articular compression fractures of the distal radius in osteoporotic women. Arch Orthop Trauma Surg 2003;123:22–27.
119. Schildhauer TA, Bauer TW, Josten C, Muhr G. Open reduction and augmentation of internal fixation with an injectable skeletal cement for the treatment of complex calcaneal fractures. J Orthop Trauma 2000;14:309–317.

See also DRUG DELIVERY SYSTEMS; MATERIALS AND DESIGN FOR ORTHOPEDIC DEVICES; POROUS MATERIALS FOR BIOLOGICAL APPLICATIONS.

BIOMATERIALS: AN OVERVIEW

BRANDON L. SEAL
ALYSSA PANITCH
Arizona State University
Tempe, Arizona

INTRODUCTION

Biomaterials are materials that are used or that have been designed for use in medical devices or in contact with the body. Traditionally, they consist of metallic, ceramic, or synthetic polymeric materials, but more recent developments in biomaterials design have attempted to incorporate materials derived from or inspired by biological materials (e.g., silk and collagen). Often, the use of biomaterials focuses on the augmentation, replacement, or restoration of diseased or damaged tissues and organs. The prevalence of biomaterials within society is most evident within medical and dental offices, pharmacies, and hospitals. However, the influence of biomaterials has reached into many households with examples ranging from increasingly common news media coverage of medical breakthroughs to the availability of custom color non-corrective contact lenses.

The evolving character of the discipline of biomaterials is evidenced by how the term biomaterial has been defined. In 1974, the Clemson Advisory Board, in response to a request by the World Health Organization (WHO), stated that a biomaterial is a "systemically pharmacologically inert substance designed for implantation within or incorporation with living tissue" (1). Dr. Jonathan Black further modified this definition to state that a biomaterial is "any pharmacologically inert material, viable or nonviable, natural product or manmade, that is part of or is capable of interacting in a beneficial way with a living organism" (1). An National Institute of Health (NIH) consensus definition appeared in 1983 and defined biomaterials as "any substance (other than a drug) or combination of substances, synthetic or natural in origin, which can be used for any period of time, as a whole or as a part of a system that treats, augments, or replaces any tissue, organ, or function of the body" (2). Thus, relatively newer definitions of the term biomaterial recognize that more modern medical and diagnostic devices will rely increasingly upon direct biological interaction between biological molecules, cells, and tissues and the materials from which these devices are manufactured.

HISTORY OF BIOMATERIALS

Compared with the much larger field of materials science, the field of biomaterials is relatively new. Although there exist recorded cases of glass eyes and metallic or wooden dental implants (some of which can be dated back to ancient Egypt), the modern age of biomaterials could not have existed without the adoption of aseptic surgical techniques pioneered by Sir Joseph Lister in the mid-nineteenth century and indeed, did not fully emerge as an industry or discipline until after the development of synthetic polymers just prior to, during, and following World War II. Prior to World War II, implanted biomaterials consisted primarily of metals (e.g., steel, used in pins and plates for bone fixation, joint replacements, and the covering of bone defects). In the late 1940s, Harold Ridley observed that shards of poly(methyl methacrylate) (PMMA), from airplane cockpit windshields, embedded within the eyes of World War II aviators did not provoke much of an inflammatory response (3). This observation led not only to the development of PMMA intraocular lenses, but also to greater experimentation of available materials, especially polymers, as biomaterials that could be placed in direct contact with living tissue.

As the fields of cellular, molecular, and developmental biology began to grow during the 1970s and 1980s, new insights into the organization, function, and properties of biological systems, tissues, and interactions led to a greater understanding of how cells respond to their environment. This wealth of biological information allowed the field of biomaterials to undergo a paradigm shift. Instead of focusing primarily on replacing an organ or a tissue with a synthetic, usually nondegradable biomaterial, a new branch of biomaterials would attempt to combine biologically active molecules, therapeutics, and motifs into existing and novel biomaterial systems derived from both synthetic and natural sources (4–6). Although there exist many examples of successful, commercially available biomaterials consisting of metallic and ceramic bases, the focus of biomaterials research has shifted to the development of polymeric or composite materials with biologically sensitive or environmentally controlled properties. This change has resulted largely due to the reactivity and variety of chemical moieties that are found in or that can be engineered into natural and synthetic polymers. Indeed, by viewing biomaterials as materials designed to interact with biology rather than being inert substances, the field of biomaterials has exploded with innovative designs that promote cell attachment, encapsulation, proliferation, differentiation, migration, and apoptosis, and that allow the biomaterial to polymerize, swell, and degrade under a variety of environmental conditions and biological stimuli. Evidence of this polymer and composite revolution is the dramatic increase in the number of publications relating to biomaterials research. Figure 1 shows a plot of the number of journal articles with biomaterial or biomaterials in their title, abstract, or keyword as a function of publication year as searched in the Web of Science database. As seen in Fig. 1, publications matching the search criteria have increased exponentially starting around the early 1990s and continuing until the present. The number of scientific journals, shown in Table 1, related to research in the field of biomaterials has also grown. Although the number of journal articles related to biomaterials research may have resulted primarily from the large increase in the number of biomaterials-related scientific journals, the exponential growth of biomaterials is evidenced further by the growth of the number of bioengineering or biomedical engineering departments (the department in which most biomaterials programs reside) established at universities throughout the United States (Fig. 1).

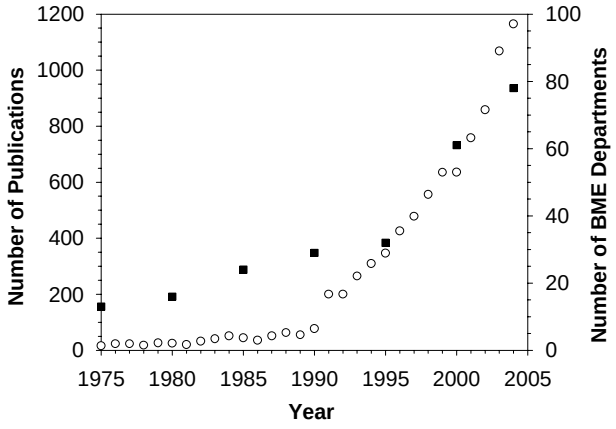


Figure 1. A plot of the number of publications (○) containing the word biomaterial or biomaterials in the title, abstract, or keywords, as searched in the Web of Science database, as a function of the publication year as well as a plot of the number of bioengineering or biomedical engineering departments (BME departments) within the United States as a function of time (■).

MARKET SIZE AND TYPES OF APPLICATIONS

The field of biomaterials, by nature, is interdisciplinary. Successful biomaterial designs have involved talents, knowledge, and expertise provided by physicians and clinicians, materials scientists, engineers, chemists, biologists, and physicists. As a result, it is not surprising that the biomaterials industry is both relatively young and very diversified. The diversity of this industry has resulted from the types of products created and marketed, the size and location of involved companies, and the types of regulatory policies imposed by government agencies and third party reimbursement organizations. Specifically, the biomaterials industry is part of the Medical Device and Diagnostic Industry, a multibillion dollar industry comprised of organizations that design, fabricate, and/or manufacture materials that are used in the health and life science fields. The end use applications are medical and dental devices, prostheses, personal hygiene products, diagnostic devices, drug delivery vehicles, and biotechnology systems. Some examples of these applications include full and hybrid artificial organs, biosensors, vascular grafts, pacemakers, catheters, insulin pumps, cochlear implants, contact lenses, intraocular lenses, artificial joints and bones, burn dressings, and sutures. Table 2 shows a list of some common medical devices that require various biomaterials, and Table 3 displays a list of the prevalence and market potential of a few of these applications (7).

GOVERNMENT REGULATION

Within the United States, in a research only environment, biomaterials by themselves do not necessarily require government regulation. However, if any biomaterial is used within a medical or diagnostic device designed and destined for commercialization, the biomaterials used within the medical device (as well as the device itself) are subject to the jurisdiction of the U.S. Food and Drug Administration (FDA) as set forth in the Federal Food, Drug, and Cosmetic Act of 1938, the Medical Device Amendments of 1976, and the Food and Drug Administration Modernization Act of 1997. These laws have empowered the FDA to regulate conditions involving premarket controls, postmarket reporting, production involving Good Manufacturing Practices, and the registration and listing of medical devices. Any biomaterial within a marketed medical device prior to the Medical Device Amendments of 1976 were grandfathered and are considered approved materials. Modifications to these materials or

Table 1. A List of Journals with Publications Related to the Field of Biomaterials[a]

Name of Journal	Name of Journal	Name of Journal
Advanced Drug Delivery Reviews (1987)	Biosensors and Bioelectronics (1985)	Journal of Biomaterials Science: Polymer Edition (1990)
American Journal of Drug Delivery (2003)	Cells and Materials (1991)	Journal of Biomedical Materials Research (1967)
American Society of Artificial Internal Organs Journal (1955)	Cell Transplantation (1992)	Journal of Controlled Release (1984)
Annals of Biomedical Engineering (1973)	Clinical Biomechanics (1986)	Journal of Drug Targeting (1993)
Annual Review of Biomedical Engineering (1999)	Colloids and Surfaces B: Biointerfaces (1993)	Journal of Long Term Effects of Medical Implants (1991)
Artificial Organs (1977)	Dental Materials (1985)	Journal of Nanobiotechnology (2003)
Artificial Organs Today (1991)	Drug Delivery (1993)	Macromolecules (1968)
Biomacromolecules (2000)	Drug Delivery Systems and Sciences (2001)	Materials in Medicine (1990)
Biofouling (1985)	Drug Delivery Technology (2001)	Medical Device and Diagnostics Industry (1996)
Biomedical Engineering OnLine (2002)	e-biomed: the Journal of Regenerative Medicine (2000)	Medical Device Research Report (1995)
Bio-medical Materials and Engineering (1991)	European Cells and Materials (2001)	Medical Device Technology (1990)
Biomaterial-Living System Interactions (1993)	Federation of American Societies for Experimental Biology Journal (1987)	Medical Plastics and Biomaterials (1994)
Biomaterials (1980)	Frontiers of Medical and Biological Engineering (1991)	Nanobiology
Biomaterials, Artificial Cells and Artificial Organs (1973)	IEEE Transactions on Biomedical Engineering (1954)	Nanotechnology (1990)
Artificial Cells, Blood Substitutes, and Immobilization Biotechnology (1973)	International Journal of Artificial Organs (1976)	Nature: Materials (2002)
Biomaterials Forum (1979)	Journal of Bioactive and Compatible	Tissue Engineering (1995)
Biomedical Microdevices (1998)	Polymers (2002) Journal of Biomaterials Applications (2001)	Trends in Biomaterials and Artificial Organs (1986)

[a]The date of first publication is listed in parentheses following the name of each journal.

new materials are subject to controls established by the FDA. These controls consist of obtaining an Investigational Device Exemption for the medical device, including the biomaterials used within the device, prior to conducting clinical trials.

In addition, biomaterials can be considered part of a Class I, II, or III device depending on FDA classifications and depending on whether or not the biomaterial is considered to be part of a biologic, drug, or medical device. Class I devices are generally considered those devices needing the least amount of regulatory control since they do not present a great risk for a patient. Examples include tongue depressors and surgical drills. Class II devices represent a moderate risk to patients and require additional regulation (e.g., mandatory performance standards, additional labeling requirements, and postmarket surveillance). Some examples include X-ray systems and cardiac mapping catheters. Class III devices (e.g., cardiovascular stents and heart valves), represent those devices with the highest risk to patients and require extensive regulatory control. Usually, for biomaterials in direct contact with tissue within the body, devices are considered Class III devices and are subject to a Premarket Approval process before they can be sold within the United States. In general, for most biomaterials, some of the tests the FDA reviews to evaluate biomaterial safety includes tests

Table 2. Some Common Uses for Biomaterials

Organ/Procedure	Associated Medical Devices
Bladder	Catheters
Bone	Bone plates, joint replacements (metallic and ceramic)
Brain	Deep brain stimulator, hydrocephalus shunt, drug eluting polymers
Cardiovascular	Polymer grafts, metallic stents, drug eluting grafts
Cosmetic enhancement	Breast implants, injectable collagen
Eye	Intraocular lenses, contact lenses
Ear	Artificial cochlea, artificial stapes
Heart	Artificial heart, ventricular assist devices, heart valves, pacemakers
Kidney	Hemodialysis instrumentation
Knee	Metallic knee replacements
Lung	Blood oxygenator
Reproductive system	Hormone replacement patches, contraceptives
Skin	Artificial skin, living skin equivalents
Surgical	Scalpels, retractors, drills
Tissue repair	Sutures, bandages

Table 3. A Summary of the Prevalence and Economic Cost of Some of the Healthcare Treatments Requiring Biomaterials for the Year 2000

Medical Application[a]	Incident Patient Population[a]	Prevalent Patient Population[a]	Total Therapy Cost (Billions of US Dollars)[a]
Dialysis	188,000	1,030,000	$67
Cardiovascular			
Bypass grafts	733,000	6,000,000	$65
Valves	245,000	2,400,000	$27
Pacemakers	670,000	5,500,000	$44
Stents	1,750,000	2,500,000	$48
Joint replacement	1,285,000	7,000,000	$41
Hips	610,000		
Knees	675,000		

[a]All data taken from that reported by Lysaght and O'Loughlin (7).

involving cellular toxicity (both direct and indirect), acute and chronic inflammation, debris and degradation byproducts and associated clearance events, carcinogenicity, mutagenicity, fatigue, creep, tribology, and corrosion. Further information regarding FDA approval for medical devices can be found on the FDA webpage, www.fda.gov.

Many FDA approved biomaterials continue to be monitored for efficacy and safety in an effort not only to protect patients, but also to improve biocompatibility and reduce material failure. Perhaps the best known example of an FDA regulated biomaterial is silicone. Silicone had been used since the early 1960s in breast implants. As a result, silicone breast implants were grandfathered into the Medical Device Amendments of 1976. During the 1980s, some concerns regarding the safety of silicone breast implants arose and prompted the FDA to request, in 1990, additional safety data from manufacturers of breast implants. Due to fears of connective tissue disease, multiple sclerosis, and other ailments resulting from ruptured silicone implants, the FDA banned silicone breast implants in 1992. Recently, however, manufacturers (e.g., the Mentor Corporation) have applied for and received premarket approval for the sale of silicone breast implants contingent on the compliance of various conditions (8). Thus, silicone is a good example of the complexity surrounding the testing of both efficacy and safety for biomaterials.

TYPES OF BIOMATERIALS

Similar to the field of materials science, the field of biomaterials focuses on four major types of materials: metals, ceramics, polymers, and composites. Examples of a few selected medical devices made from these materials are shown in Fig. 2. The materials selected for any particular application depend on the properties desired for a particular function or set of functions. In all materials applications, the structure, properties, and processing of the selected material will affect performance. As a result, physicians, scientists and engineers who design biomaterials need to understand not only mechanical and physical properties of materials, but also biological properties of materials. Mechanical and physical properties include strength, fatigue, creep resistance, flexibility, permeability

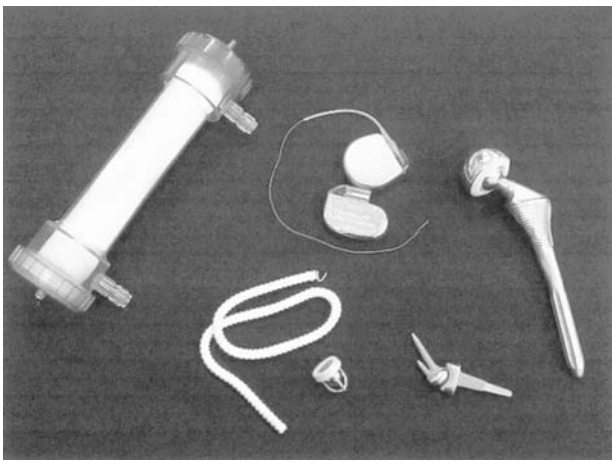


Figure 2. A representation of a few medical devices made from various biomaterials. From the upper left corner and moving clockwise, this picture shows a kidney hemodialyzer, two pacemakers, a hip replacement, an articulating wrist joint, a heart valve, and a vascular graft.

to gases and liquids, thermal and electrical properties, chemical reactivity, and degradation. Biological properties of materials largely focus on biocompatibility issues related to toxicity, immune system reactivity, thrombus formation, tribiology, inflammation, carcinogenic and teratogenic potential, integration with tissues and cells, and the ability to be sterilized. Regardless of the material, recent approaches to biomaterials research has focused on directing specific tissue interaction by using materials to introduce chemical bonds with the surrounding tissue, to act as scaffolds for tissue ingrowth, to introduce an inductive signal that will influence the behavior of surrounding cells or matrix, or to form new tissue when incubated or presented to transplanted cells.

Metals

Metals have been used as biomaterials for centuries. Although some fields (e.g., dentistry) continue to use amalgams, gold, and silver, most modern metallic biomaterials consist of iron, cobalt, titanium, or platinum bases. Since they are strong, metals are most often employed as biomaterials in orthopedic or fracture fixation medical devices; however, metals are also excellent conductors, and are therefore used for electrical stimulation of the heart, brain, nerves, muscle, and spinal cord. The most common alloys for orthopedic applications include stainless steel, cobalt, and titanium alloys. These alloys have enjoyed frequent use in medical procedures related to the function of joints and load-bearing. For example, metal alloys are commonly found in medical devices for knee replacement as well as in the femoral stem used in total hip replacements. Since all metals are subject to corrosion, especially in the salty, aqueous environment within the body, metals used as biomaterials often require an external oxide layer to protect against pitting and corrosion. These electrochemically inert oxide layers consist of Cr_2O_3 for stainless steel, Cr_2O_3 for cobalt alloys, and TiO_2 for

Figure 3. Photographs of stainless steel (a), cobalt–chromium (b), and titanium alloy (Ti6Al4V) (c) hip implants. (All three photographs are used with permission from the Department of Materials at Queen Mary University of London.)

titanium alloys. Figure 3 displays examples of three types of metallic hip replacements.

Stainless Steel Alloys. The stainless steel most commonly used as orthopedic biomaterials is classified 316L by the American Iron and Steel Institute. This particular austenitic alloy contains a very low carbon content (a maximum of 0.03%) and chromium content of 17–20%. The added chromium will react with oxygen to produce a corrosion-resistant chromium oxide layer. The 316L grade of stainless steel is a casting alloy, and its relatively high ductility makes this alloy amenable to extensive postcasting mechanical processing. Compared to cobalt and titanium alloys, stainless steel has a moderate yield and ultimate strength, but high ductility. Furthermore, it may be fabricated by virtually all machining and finishing processes and is generally the least expensive of the three major metallic alloys (4,5,9).

Cobalt Alloys. Cobalt alloys have been used since the early twentieth century as dental alloys and in heavily loaded joint applications. For use as a biomaterial, cobalt alloys are either cast (i.e., primarily formed within a mold) or wrought (i.e., worked into a final form from a large ingot). Two examples of cobalt alloys include Vitallium (designated F 75 by ASTM International), a cast alloy that consists of 27–30% chromium and >34% cobalt, and the wrought cobalt alloy MP35N (designated F 563 by ASTM International), which consists of 18–22% chromium, 15–25% nickel, and >34% cobalt. Compared to Vitallium, the MP35N alloy has demonstrated superior fatigue resistance, larger ultimate tensile strength, and a higher degree of corrosion resistance to chlorine. Consequently, this particular alloy is good for applications requiring long service life without fracture or stress fatigue. Compared to stainless steel alloys, cobalt-based alloys have slightly higher tensile moduli, but lower ductility. In addition, they are more expensive to manufacture and more difficult to machine. However, relative to stainless steel and titanium, cobalt-based alloys can offer the most useful balance of corrosion resistance, fatigue resistance, and strength (4,5,9).

Titanium Alloys. The most recent of the major orthopedic metallic alloys to be employed as biomaterials are titanium alloys. Although pure titanium is relatively weak and ductile, titanium can be stabilized by adding elements (e.g., aluminum and vanadium) to the alloy. Often, pure titanium (designated F 67 by ASTM International) is primarily used as a surface coating for orthopedic medical devices. For load-bearing applications, the alloy Ti6Al4V (designated F 136 by ASTM International) is much more widely used in implant manufacturing. As in the case of stainless steel and cobalt alloys, titanium contains an outer oxide layer, composed of TiO_2, that protects the implant from corrosion. In fact, of the three major orthopedic alloys, titanium shows the lowest rate of corrosion. Moreover, the density of titanium is almost half that of stainless steel and cobalt alloys. As a result, implants made from titanium are lighter and reduce patient awareness of the implant; however, titanium alloys are among the most expensive metallic biomaterials. Relative to stainless steel and cobalt alloys, titanium has a lower Young's modulus, which can aid in reducing the stresses around the implant by flexing with the bone. Titanium has a lower ductility than the other alloys, but does demonstrate high strength. These properties allow titanium alloys to play a diverse role as a biomaterial. Titanium alloys are used in parts for total joint replacements, screws, nails, pacemaker cases, and leads for implantable electrical stimulators (4,5,9).

Despite the reduced weight and improved mechanical match of titanium alloy implants to bone relative to stainless steel and chromium alloy implants, titanium alloy implants still exhibit issues with regard to mechanical mismatch. This problem stems from the large differences in properties (e.g., elastic moduli) between bone, metals, and polymers used as acetabular cups. For example, metals have elastic moduli ranging from ~100 to 200 GPa, ultrahigh molecular weight polyethylene has an elastic modulus of 1–2 GPa, and the elastic modulus of cortical bone is ~ 12 GPa (10). In addition, it is difficult to produce a titanium implant surface that is conducive to bone ingrowth or attachment. Novel titanium foams have been investigated as a method for reducing implant weight, better matching tissue mechanics, and improving bone ingrowth. The process involves mixing titanium powder with ammonium hydrogen carbonate powder and compressing and heating the mixture to form foams with densities varying from 0.2 to 0.65 times the density of solid titanium. These densities are close to those of cancellous bone (0.2–0.3 times the density of solid titanium) and cortical bone (0.5–0.65 times the density of solid titanium) (11). While they are preliminary, studies with novel materials such as these titanium foams illustrate a trend toward the development of materials that better mimic the properties of the native tissue they are designed to replace.

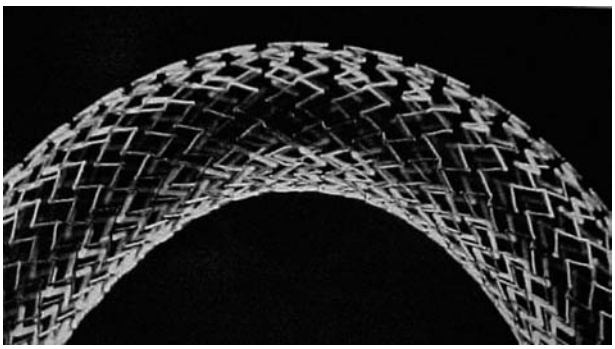


Figure 4. A photograph of the SMARTeR nitinol stent developed by the Cordis Corporation. (Reprinted from Ref. 12, with permission from Royal College of Radiologists.)

Other Metals. Besides stainless steel, cobalt alloys, and titanium alloys, there exist other examples of metals used as biomaterials. Some examples include nitinol, a single-phase nickel/titanium shape memory alloy, tantalum, a very dense, chemically inert, weak, but fatigue-resistant metal, and platinum, a very expensive metal used by itself or with iridum as a corrosion-resistant electrical conductor for electrode applications. Nitinol stents (e.g., that seen in Fig. 4) (12) and drug-eluting nitinol stents used for cardiovascular applications recently have seen enormous medical and commercial success. Indeed, metallic stents have significantly changed the way coronary blockages are treated (4,5,9).

Ceramics

Of the major types of materials used as biomaterials, ceramics have not been used as frequently as metals, polymers, or composites. However, ceramics continue to enjoy widespread use in certain bone-related applications (e.g., dentistry and joint replacement surgeries), due to their high compressive strength, high degree of hardness, excellent biocompatibility, superior tribological properties, and chemical inertness. Although they are very strong in compression, ceramics are susceptible to mechanical and thermal loading, have lower tensile strengths relative to other materials, and are very brittle in tension; this brittleness limits potential biomaterials applications.

Ceramics consist of a network of metal and nonmetal ions, with the general structure X_mY_n, arranged in a repeating structure. This structure depends on the relative size of the ions as well as the number of counterions needed to balance total charge. For example, if $m = n = 1$, and both ions are approximately the same size, then the structure would be of a simple cubic nature (e.g., CsCl or CsI); if the anion is much larger than the cation, then typically, a face centered cubic (fcc) structure would emerge (e.g., ZnS or CdS). If $m = 2$ and $n = 3$, as is the case with oxide ceramics (e.g., Al_2O_3), then a hexagonal closed pack structure would often result (13).

Ceramics used as biomaterials can be classified by processing–manufacturing methods, by chemical reactivity, or by ionic composition. Regarding chemical reactivity, ceramics can be bioinert, bioactive, or bioresorbable. Bioinert or nonresorbable ceramics are either porous or nonporous and are essentially not affected by the environment at the implant site. Bioactive or reactive ceramics are designed with specific surface properties that are intended to react with the local host environment and to elicit a desired tissue response. Bioresorbable ceramics dissolve over some prescribed period of time *in vivo* mediated by physiochemical processes. If one considers the application of bone replacement, then there would be about four ways for ceramics to interact with and attach to bone. First, a nonporous, inert ceramic material could be attached via glues, surface irregularities, or press-filling methods. Second, a porous, inert ceramic could be designed to have an optimal pore size, which promotes direct mechanical attachment of bone through bone ingrowth. Third, a nonporous, inert ceramic with a reactive surface could direct bone attachment via chemical bonding. Fourth, a nonporous or porous, resorbable ceramic could eventually be replaced by bone. When describing real examples of ceramics used as biomaterials, it is more useful to classify the ceramics based on ionic composition. This type of classification reveals a few major bioceramic groups: oxide ceramics, multiple oxides of calcium and phosphorus, glasses and glass ceramics, and carbon.

Oxide Ceramics. As their name implies, oxide ceramics consist of oxygen bound to a metallic species. Oxide ceramics are chemically inert, but can be nonporous or porous. One example of a nonporous oxide ceramic used as a biomaterial is aluminum oxide, Al_2O_3. Highly pure aluminum oxide (F 603 as designated by ASTM International), or alumina, has high corrosion resistance, good biocompatibility, high wear resistance, and good mechanical properties due to high density and small grain size. Aluminum oxide has been manufactured as an acetabular cup for total hip replacement. In comparison with metal or ultrahigh molecular weight polyethylene, Al_2O_3 provides better tribological properties by greatly decreasing friction within the joint and substantially increasing wear resistance. Recently, the FDA approved ceramic on ceramic hip replacements made from alumina and marketed by companies such as Wright Medical Technology and Stryker Osteonics. This ceramic on ceramic design is very resistant to wear and results in a much smaller amount of wear debris than traditional metal–polymer joints. With better wear properties and longer useful lifespan, ceramic on ceramic hip replacements likely will provide an attractive alternative to other biomaterial options, especially for younger patients that need better long-term solutions for joint replacements (4,5,9).

Ceramic oxides can also be porous. In bone formation, these pores are useful for allowing bone ingrowth, which will stabilize the mechanical properties of the implant without sacrificing the chemical inertness of the ceramic material. In general, there are three ways to make a porous ceramic oxide. First, a soluble metal or salt can be mixed with the ceramic and etched away. Second, a foaming agent that evolves gases during heating (e.g., calcium carbonate) can be mixed with the ceramic powder prior to firing. Third, the microstructure of corals can be used as a template to create a ceramic with a high degree

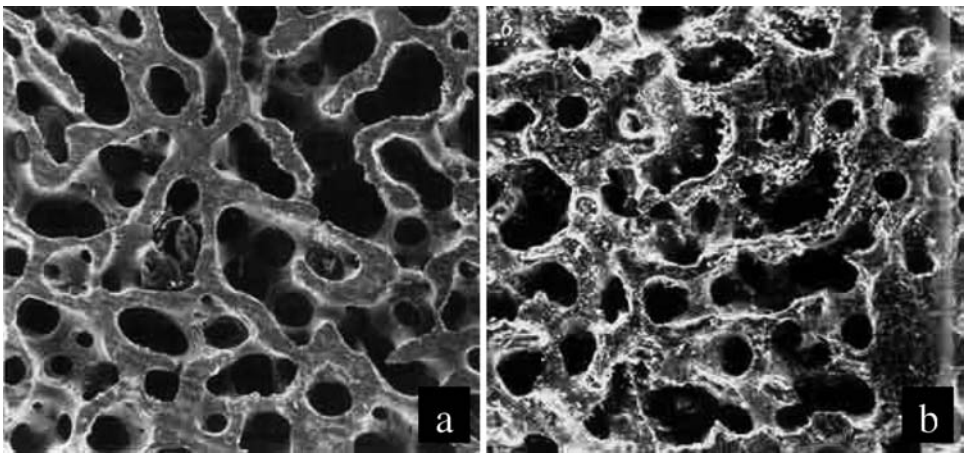


Figure 5. Photographs of (a) a cross-section of human cancellous bone and (b) coral of the genus Porites. These images illustrate how biologically derived materials (e.g., coral) can be used as scaffolds to create ceramic biomaterials that mimic the structure and porosity of natural bone. (Both photographs are used with permission from Biocoral, Inc.)

of interconnectivity and uniform pore size. In this third approach, coral is machined into the desired shape. Then, the coral is heated up to drive off carbon dioxide. The remaining calcium oxide provides a scaffold around which the ceramic material is deposited. After firing, the calcium oxide can be dissolved using hydrochloric acid. This dissolved calcium oxide will leave behind a very uniform and highly interconnected porous structure. Interestingly, the type of coral used will affect the pore size of the resulting ceramic. For example, if the genus *Porites* is used, then the pore size will range from 140 to 160 μm; the genus *Goniopora* will result in a pore size of 200–1000 μm (5). Porous ceramics do have many advantages for bone ingrowth, especially since the porous structure more closely mimics that of cancellous bone (see Fig. 5). However, the porous structure does result in a loss of strength and a tremendous increase in surface area that interacts with an *in vivo* saline environment.

Multiple Oxides of Calcium and Phosphorus. Aside from many types of proteins, the extracellular environment of bone contains a large concentration of organic mineral deposits known as hydroxyapatite. Chemically, hydroxyapatite generally has the following composition: $Ca_{10}(PO_4)_6(OH)_2$. Since hydroxyapatite is a naturally occurring ceramic produced by osteoblasts, it seemed reasonable to apply hydroxyapatite as filler or as a coating to allow better integration with existing bone. Coatings of hydroxyapatite have been applied (usually by plasma spraying) to metallic implants used in applications requiring bone ingrowth to provide a tight fit between bone and the implanted device, to minimize loosening over time, and to provide some measure of isolation from the foreign body response. Although hydroxyapatite is the most commonly used bioceramic containing calcium and phosphorus, there do exist other forms of calcium and phosphorus oxides including tricalcium phosphate, $Ca_3(PO_4)_2$, and octacalcium phosphate, $Ca_8H_2PO_4{\cdot}5H_2O$ (4,5,9,14).

Glasses and Glass Ceramics. Just as in the case of traditional glass, glass ceramics used as biomaterials contain large amounts of silica, SiO_2. Glass ceramics are formed using controlled crystallization techniques during which silica is cooled down at rate slow enough to allow the formation of a hexagonal crystal structure with small, crystalline grains (~1 μm) surrounded by an amorphous phase. Bioactive glass ceramics have been studied as biomaterials because they can attach directly to tissue via chemical bonds, they have a low thermal coefficient of expansion, they have good compressive mechanical strength, the mechanical strength of the glass–tissue interface is close to that of tissue, and they resist scratching and abrasion. Unfortunately, as with all ceramics, bioactive glasses are very brittle. Two well-known examples of commercially available glass ceramics include Bioglass, which consists of SiO_2, Na_2O, CaO, and P_2O_5, and Ceravital, which contains SiO_2, Na_2O, CaO, P_2O_5, K_2O, and MgO. Relative to traditional soda lime glass, bioactive glass ceramics contain lower amounts of SiO_2 and higher amounts of Na_2O and CaO. The high ratio of CaO to P_2O_5 in bioactive ceramics allows the rapid formation of a hydroxycarbonate apatite (HCA) layer at alkaline pH. For example, a 50 nm layer of HCA can form from Bioglass 45S5 after 1 h. The release of calcium, phosphorus, and sodium ions from bioactive ceramics also allows the formation of a water-rich gel near the ceramic surface. This cationic-rich environment creates a locally alkaline pH that helps to form HCA layers and provide areas of adhesion for biological molecules and cells (4,5,9).

Carbon. Processed carbon has been used in biomaterials applications as a bioceramic coating. Although carbon can exist in several forms (e.g., graphite, diamond), bioceramic carbons consist primarily of low temperature isotropic (LTI) and ultralow temperature isotropic (ULTI) carbon. This form of carbon is synthesized through the pyrolysis of hydrocarbon gases resulting in the deposition of isotropic carbon in a layer ~ 4 mm thick. Advantages to LTI and ULTI carbon include high strength, an elastic modulus close to that of bone, resistance to fatigue compared with other materials, excellent resistance to thrombosis, superior tribological properties, and excellent bond strength with metallic substances. The LTI carbon has been used as a coating for heart valves; however, applications remain limited primarily to coatings due to processing methods (4,5,9).

Polymers

Since the early to mid-twentieth century, the discovery of organic polymerization schemes and the advent of new polymeric species have fueled an incredible interest in the research of biomaterials. The popularity of polymers as potential biomaterials likely stems from the fact that polymers exist in a seemingly endless variety, can be easily fabricated into many forms, can be chemically modified or synthesized with chemically reactive moieties that interact with biological molecules or living tissues and cells, and can have physical properties that resemble that of natural tissues. Some disadvantages to polymeric biomaterials include relatively low moduli, instability following certain forms of sterilization, lot-to-lot variability, a lack of well-defined standards related to manufacturing, processing, and evaluating, and, for some polymers, hydrolytic instability, the need to add potentially toxic polymerization catalysts, and tissue biocompatibility of both the polymer and potential degradation byproducts. There also exist some characteristics of polymers that can be advantageous or disadvantageous depending on the application and type of polymer. Some of these characteristics include polymer degradation, chemical reactivity, polymer crystallinity, and viscoelastic behavior. Early examples of polymeric biomaterials included nylon for sutures and cellulose for kidney dialysis membranes, but more recent developments in the design of polymeric biomaterials are leading the field of biomaterials to embrace cellular and tissue interactions in order to directly induce tissue repair or regeneration.

Polymers consist of an organic backbone from which other pendant molecules extend. As their name implies, polymers consist of repeating units of one or more "mers". For example, polyethylene consists of repeating units of ethylene; nylon is comprised of repeating units of a diamine and a diacid. In general, polymers used as biomaterials are made in one of two ways: condensation or addition reactions. In condensation reactions, two precursors are combined to form larger molecules by eliminating a small molecule (e.g., water). Examples of condensation polymeric biomaterials include nylon, poly(ethylene terephthalate) (Dacron), poly(lactic acid), poly(glycolic acid), and polyurethane. In addition to synthetic polymers, biological polymers (e.g., cellulose and proteins) are formed through condensation-like polymerization mechanisms. The other major polymerization mechanism used to synthesize polymers is addition polymerization. In addition polymerization, an initiator or catalyst (e.g., free radical, heat, light, or certain ions) is used to promote a rapid polymerization reaction involving unsaturated bonds. Unlike condensation reactions, addition polymerization does not result in small molecular byproducts. Furthermore, polymers can be formed using only one type of monomer or a combination of several monomers susceptible to free radical initiation and propagation. Some examples of addition reaction polymeric biomaterials include polyethylene, poly(ethylene glycol) (PEG), poly(*N*-isopropylacrlyamide), and poly(hydroxyethyl methacrylate) (PHEMA). The chemical structure of various synthetic and natural polymers used as biomaterials are shown in Figs. 6a and b (15).

The properties of polymers are affected greatly by chemical composition and molecular weight. In general, as polymer chains become longer, their mobility decreases, but their strength and thermal stability increases. The tacticity and size of pendant chains off the backbone will affect temperature-dependent physical properties. For example, small side groups that are regularly oriented in an isotactic or syndiotactic arrangement will allow the polymer to crystallize much more readily than a polymer containing an atactic arrangement of bulky side groups. The crystalline and glass transition temperatures of polymers will affect properties (e.g., stiffness, mechanical moduli, and thermal stability) *in vivo* and will consequently influence the potential application and utility of the polymer system as a biomaterial. When the functionality of a monomer exceeds two, then the polymer will become branched upon polymerization. If a sufficient number of these high functionality monomers exist within the material, then the main chains of the polymer will become chemically cross-linked. Cross-linked polymers can be much stronger and more rigid than noncross-linked polymers. However, like linear and branched polymers, cross-linked polymers can be designed such that they degrade through hydrolytic or enzymatic mechanisms.

Due to their weaker moduli compared with that of metals or ceramics, polymers are not often used in load-bearing biomaterial applications. One exception to this observation is the example of ultrahigh molecular weight polyethylene (UHMWPE), which has a molecular weight $\sim 2{,}000{,}000\ \mathrm{g \cdot mol^{-1}}$ and has a higher modulus of elasticity than high or low density polyethylene. Additionally, UHMWPE is tough and ductile and demonstrates good wear properties and low friction. As a result, UHMWPE has been used extensively in the manufacturing of acetabular cups for total hip replacements. As an acetabular cup, UHMWPE is used in conjunction with metallic femoral stems to act as a load-bearing, low wear and friction interface. Some drawbacks to using UHMWPE include water absorption, cyclic fatigue, and a somewhat significant creep rate (4,5,9). Part of the problems surrounding UHMWPE involves its lower elastic modulus ($\sim$1–2 GPa) relative to bone ($\sim$12 GPa) and metallic implants ($\sim$100–200 GPa)

Polymers in Sutures. One of the first widespread uses of polymers as biomaterials involved sutures. In particular, polyamides and polyesters are among the most common suture materials. Nylons, an example of a polyamide, have an increased fiber strength due to a high degree of crystallinity resulting from interchain hydrogen bonding between atoms of the amide group. Nylon can be attacked by proteolytic enzymes *in vivo* and can absorb water. As a result, nylon has been used more as a short-term biomaterial. Polyester sutures, such as poly(glycolic acid), poly(lactic acid), and poly(lactic-*co*-glycolic acid) are readily degraded through hydrolytic mechanisms *in vivo*. Since one side chain of lactic acid contains a bulky hydrophobic methyl group (relative to the hydrogen side group of glycolic acid), polyesters comprised principally of lactic acid degrade at a rate slower than that of polyesters consisting mostly of glycolic acid. The degradation rate of copolymers

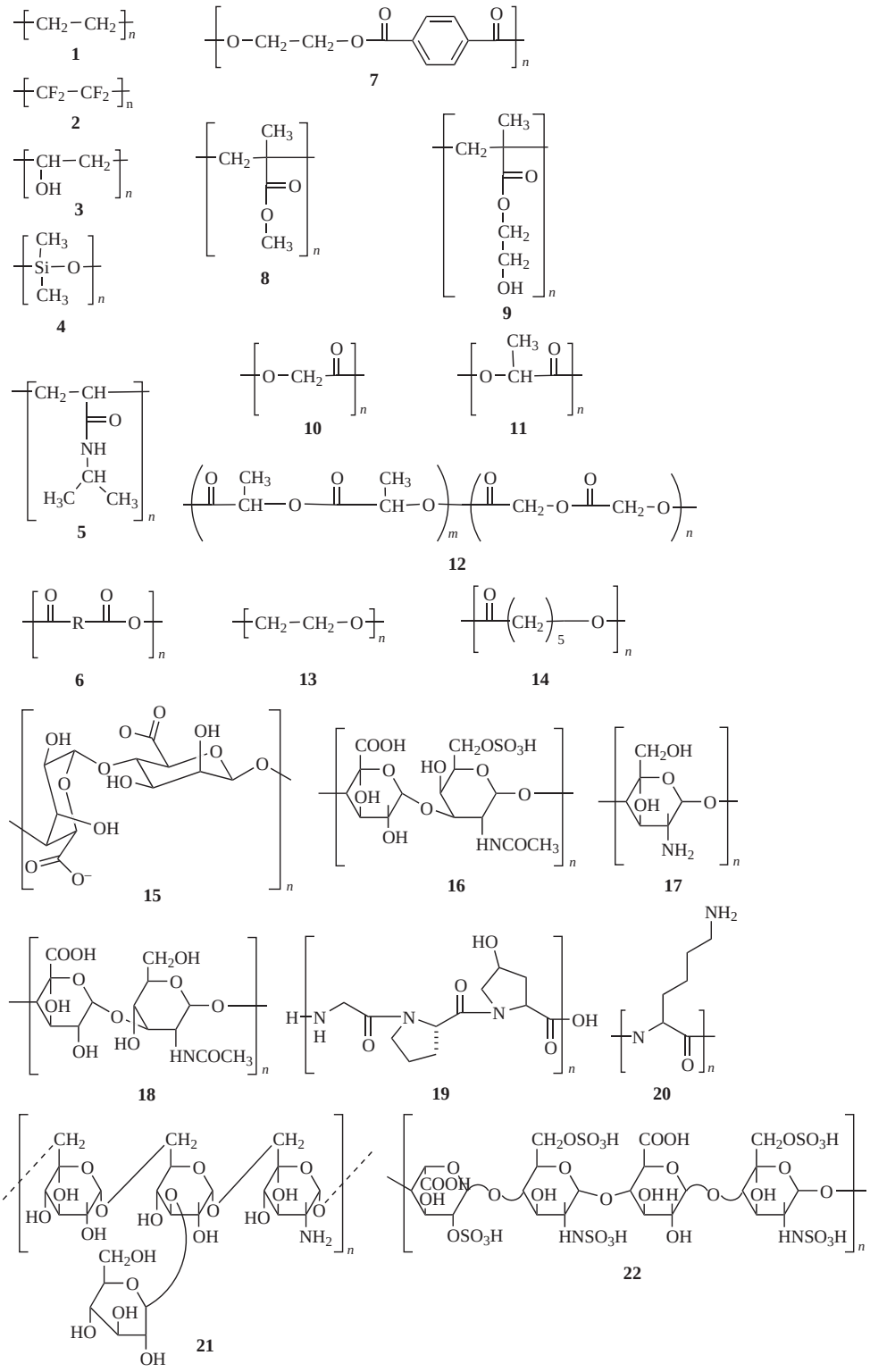


Figure 6. (a) Chemical schematics representing synthetic polymers used as biomaterials. The structures represent polyethylene (**1**), polytetrafluoroethylene (**2**), poly(vinyl alcohol) (**3**), poly(dimethyl siloxane) (**4**), poly(N-isopropylacrylamide) (**5**), polyanhydride (**6**), poly(ethylene terephthalate) (**7**), poly(methyl methacrylate) (**8**), poly(hydroxyethyl methacrylate) (**9**), poly(glycolic acid) (**10**), poly(lactic acid) (**11**), poly(lactic-co-glycolic acid) (**12**), poly(ethylene oxide) (**13**), and poly(e-caprolactone) (**14**). (Adopted from Ref. 15 with permission from Elsevier.) (b) Chemical schematics representing naturally derived polymers used as biomaterials. The structures represent alginate (**15**), chondroitin-6-sulfate (**16**), chitosan (**17**), hyaluronan (**18**), collagen (**19**), polylysine (**20**), dextran (**21**), and heparin (**22**). (Reprinted from Ref. 15 with permission from Elsevier.)

Figure 7. A photograph of a Carboflo vascular graft made out of expanded polytetrafluoroethylene (ePTFE) impregnated with carbon and marketed by Bard Peripheral Vascular, Inc.

of glycolic and lactic acid can be tailored based on the relative molar ratios of each monomer. Although the local pH of degrading polyesters can cause local inflammation concerns, the degradation byproducts of glycolic and lactic acid can be readily cleared through existing biochemical pathways. As a result, polyester sutures are commonly used within the body in applications where removal of sutures would warrant an invasive procedure (4,5,16).

Polymers in Cardiovascular Applications. Poly(ethylene terephalate) (Dacron) and expanded polytetrafluoroethylene (Teflon) have been used for decades as vascular grafts. An example of a Teflon vascular graft is shown in Fig. 7. Both of these polymers have excellent burst strengths and can be sutured directly to existing vasculature. For applications involving large diameter vascular grafts (> 6 mm), these two materials have worked well. However, neointimal hyperplasia and thrombus formation severely limit the patency of all known polymeric materials used for small diameter vascular grafts (17). Most current strategies to improve vascular graft patency involves chemically modifying the polymers used as vascular grafts to include the anticoagulant heparin, endothelial binding peptide analogues, and growth factors to stimulate endothelialization and minimize proliferation of smooth muscle into the lumen of the graft (15,18).

Polymers for Tissue Engineering. For many *in vivo* applications, researchers continue to evaluate a variety of polymeric biomaterials. Some more recent additions to the repertoire of biomaterials include naturally derived or recombinantly produced biological polymers. As an example, in the case of articular cartilage repair, it is evident that many types of polymers can be designed, modified, or combined with other materials to create new generations of biomaterials that promote healing and/or restore biological function. For example, synthetic polymers, such as poly(vinyl alcohol) (PVA), PMMA, poly(hydroxyethyl methacrylate), poly(*N*-isopropylacrylamide), polyethylene, poly(lactic acid), poly(glycolic acid), poly(lactic-*co*-glycolic acid), and poly(ethylene glycol) and naturally derived polymers (e.g., alginate, agarose, chitosan, hyaluornic acid, collagen, and fibrin) have been studied extensively with and without biochemical modifications to replace cartilage function or to promote neocartilage formation (15,19,20). These and other polymeric biomaterials have been used in studies related to liver, nerve, cardiovascular, bone, ophthalmic, skin, and pancreatic repair or restoration (15,21).

Hydrogels. As the name implies, hydrogels are polymer networks that contain large amounts of water (up to or $> 90\%$ water). As a result, hydrogels generally are hydrophilic materials, although, the presence of hydrophobic domains within the hydrogel backbone can enhance mechanical properties. To avoid dissolution into the aqueous phase, the polymeric component of the hydrogel must contain cross-links. The majority of hydrogel systems use chemical cross-links, such as covalent bonds to create a three-dimensional (3D) network; however, some hydrogels exist that rely on physical interactions to maintain gel integrity.

The high water content of hydrogels provides many benefits. First, of all the materials within materials science, the physical and mechanical properties of hydrogels most closely resemble those of biological tissue. Due to their polymeric content, hydrogels exhibit viscoelastic behavior. The elastic modulus, G', of many gel compositions reaches 1 MPa, but some hydrogels can be as strong as 20 MPa. These mechanical properties match well with those reported for many tissues. Second, the large presence of water within hydrogels can limit nonspecific interactions within the body, can shield the polymer from leukocytes and can decrease frictional effects at the site of implantation. Third, the relatively low concentration of polymer within the hydrogel can result in materials with higher porosities. Consequently, it is possible not only for cells to migrate within the hydrogel structure, but also for nutrients and waste products to diffuse into and out of the gel structure (15,22,23).

In addition to high water content, hydrogels possess other characteristics that are beneficial for biomedical applications. For example, chemical composition of polymers used in hydrogel formulations is amenable to chemical modification of the backbone and/or side group structures. These polymer derivatives allow the incorporation of various gelation chemistries, degradation rates and biologically active molecules. Although not a complete list, some of the polymers used as biomaterial hydrogels include poly(ethylene glycol), PVA, poly(hydroxyethyl methacrylate) PHEMA, poly(*N*-isopropylacrylamide), poly(vinyl pyrrolidone), dextran, alginate, chitosan, and collagen. These hydrogels, in addition to many others, are currently being explored as materials for use in cartilage, skin, liver, nerve, muscle, cardiovascular, and bone tissue engineering applications.

Poly(ethylene glycol). One of the most widely studied hydrogel materials is PEG, which contains repeats of the monomer CH_2CH_2O and exhibits a large radius of hydration due to its high hydrophilicity. As a result, PEG

can avoid detection by the body, and often is coupled to pharmaceuticals or other molecules to extend circulation half-life within the body. Of all the materials used in biomedical research, few polymers have better biocompatibility properties than PEG. Also, the chemical structure of PEG is fairly stable within aqueous environments, although hydrolytic degradation can occur. Furthermore, removal of PEG from the body is not a major concern since PEG, with a molecular weight $< 20{,}000\,g{\cdot}mol^{-1}$, can be cleared readily by the kidneys. Traditionally, PEG hydrogels have been cross-linked through chemical initiators, however, other work has shown that photoinitiators can be used to gel PEG *in situ*. Recently, more attention has focused on the use of star PEG, which contain a central core out of which proceeds several linear PEG arms. Consequently, these materials offer improved control over mechanical properties and biological interactions since each molecule of polymer contains many more potential sites for cross-linking or for incorporating biologically active molecules. Cell adhesion peptides, polysaccharides, and polysaccharide ligands have all been coupled to various PEG molecules and studied as biomaterials (15,23,24).

Acrylics. One of the greatest success stories involving polymeric biomaterials involves PMMA and PHEMA. Many polymers have not yet been approved by the FDA. However, many polymers of the acrylic family (e.g., PMMA used for bone cement and intraocular lenses) were grandfathered into the Medical Device Amendments of 1976 as approved materials. The PHEMA polymer allows for sufficient gas exchange, and both PHEMA and PMMA have excellent optical properties and a good degree of hydration. As a result, intraocular lenses, hard, and soft contact lenses (see Fig. 8) made in whole or in part from these polymers are commercially available (3,5). Even though contact lenses only touch the eye on one side, the polymers that comprise the contact lenses are still bathed in tears and are therefore subject to protein deposition. This protein deposition can cause eye irritation and lead to contact lens failure if the contacts are not properly cleaned. With the development of disposable contact lenses, however, the problem of protein buildup can be minimized since the useful lifespan of each contact lens does not have to extend very long.

Biomimetic Materials. Recently, the field of biomaterials has started to incorporate features found within mechanisms involved in biomolecular assembly and interaction (24). These biomimetic materials show great promise since assembly is directed through biological affinity, recognition and/or interactions. As a result, these materials often have properties more similar to those of natural materials. Biomimetic materials exist as polymer scaffolds and hydrogels, but can also consist of ceramic and metallic materials machined or chemically modified to mimic porous structure of tissue (e.g., bone). Further elucidation of mechanisms responsible for biological self-assembly most likely will lead to improved biomaterials that are capable of interacting very specifically with an environment containing cells, tissue, or ECM molecules. In addition, many researchers are borrowing biological concepts to provide appropriate signals for cellular proliferation or differentiation and to deliver pharmaceuticals in a much more controlled manner. The scanning electron micrograph (SEM) shown in Fig. 9 illustrates a biologically oriented approach of using a biomaterial like chitosan–collagen as a scaffold on which cells can adhere (25).

Drug Delivery. Applications involving biomaterials have evolved from those focused on mostly structural requirements to those combining multiple design considerations including structure, mechanics, degradation, and drug delivery. The latest trend in biomaterials design is to promote healing, repair, or regeneration via the delivery of pharmaceutical agents, drugs, or growth factors. There exist many examples of biomaterials used as delivery vehicles or as drugs (22); however, many of these examples

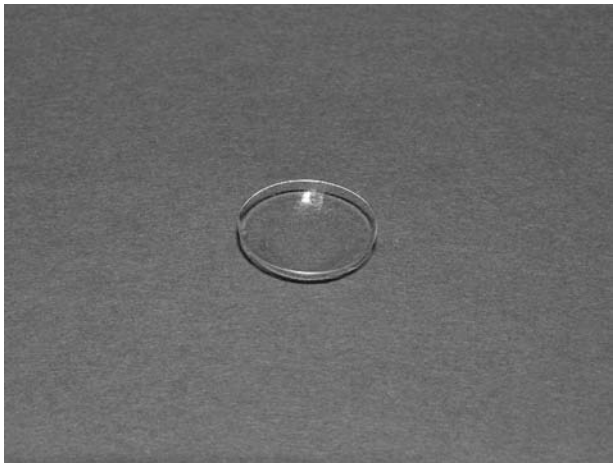


Figure 8. A photograph of a disposable contact lens made from PHEMA.

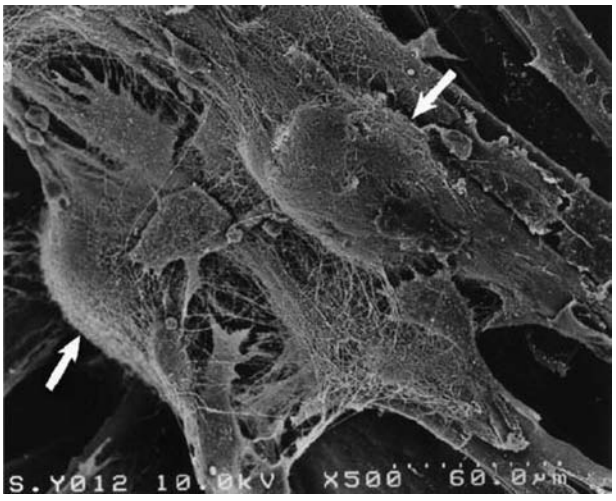


Figure 9. A SEM showing chondrocytes (denoted by white arrows) attached to a biomaterial scaffold comprised of chitosan-based hyaluoronic acid hybrid polymer fibers. (Reprinted from Ref. 25 with permission from Elsevier.)

are just beginning to transition from research materials into commercially available products. One example of a commercially available drug delivery biomaterial is known as Gliadel, made by Guilford Pharmaceuticals. Gliadel wafers consist of a polyanhydride polymer loaded with carmustine, a chemotherapeutic drug. This system is intended as a treatment for malignant gliomas. Following removal of the tumor, the Gliadel wafers are added to the cavity and allowed to degrade and release the carmustine in order to kill remaining tumor cells. In addition to cancer treatments, biomaterials as drug delivery vehicles have been extensively employed in cardiovascular applications. Recently, FDA approval was granted to several types of drug eluting metallic stents. Among these include the Sirolimus-eluting CYPHER stent manufactured by the Cordis Corporation, the TAXUS Express^2 Paclitaxel-eluting stent manufactured by the Boston Scientific Corporation. The purpose behind releasing the drugs from the stents is to decrease the occurrence of restenosis, or the renarrowing of vessels treated by the stent. As a result of the drug delivery aspect of the system, the stents are expected to have better long-term viability. Several more examples of drug delivery and biomaterial hybrid systems exist; however, a comprehensive review of biomaterials as drug delivery systems is beyond the scope of this article. It is important to note that more interest and attention have been given to modify biomaterials so that the material is more integrally involved in interacting with and manipulating organ and tissue biology.

FACTORS CONTRIBUTING TO BIOMATERIAL FAILURE

Although there exists a multitude of commercially available and successful metallic, ceramic, and polymeric biomaterials, biomaterials have and will continue to fail. The human body is a very hostile environment for synthetic and natural materials. In some instances, like orthopedic applications, it is much easier to understand why materials can fail since no material can survive cyclical loading indefinitely without showing signs of fatigue or wear. However, for most biomaterial failures, the exact reason for failure is still not well understood. Some factors contributing to the failure of a biomaterial include corrosion, wear, degradation, and biological interactions.

Corrosion

By weight, more than one-half of the human body consists of water. As a result, all implanted biomaterials will encounter an aqueous environment. Moreover, this aqueous environment is also very saline due to the presence of a relatively large concentration of extracellular salts. The aqueous and saline conditions of physiological solutions create favorable conditions for metallic corrosion. Corrosion involves oxidation and reduction reactions between a metal, ions, and species (e.g., dissolved oxygen). In fact, the lowest free energy state of many metals in and oxygenated and hydrated environment is an oxide. Most corrosion reactions are electrochemical. For example, if zinc metal is placed in an acidic environment (e.g., hydrochloric acid), hydrogen gas will evolve as the zinc become cationic and binds to chloride ions. The actual reaction consists of two half reactions. In the first reaction, zinc metal is oxidized to a Zn^{2+} state; the second reaction involves the reduction of hydrogen ions to hydrogen gas. During this process, the newly formed metal ions diffuse into solution. Both the oxidation and reduction reactions must occur at the same time to avoid charge buildup within the material. This process occurs at the surface and exposed pore of metals, and, in an attempt to passivate the surface to avoid this process, corrosion resistant oxides have been incorporated into an implant surface (13). Care must be taken, however, to ensure that the protective oxide coating is not damaged during processing, packaging, or surgical procedure.

In addition to oxidative corrosion, bimetallic or galvanic corrosion is a concern with implants composed of more than one type of metal, such as alloys with mixing defects and implants containing parts made from distinct metals. Galvanic corrosion can occur because all metals have a different tendency to corrode. If two distinct metals are in contact with one another through a conductive medium, oxidation of one metal will occur while reduction of the other occurs. In both oxidative corrosion and bimetallic corrosion, bits of metal, metal ions, and oxidative debris can enter the surrounding tissue and even travel to distant body parts. This can result in inflammation and even in metal toxicity.

Wear

In addition to corrosion, metal, as well as other materials can wear as a result of friction. For example, in hip implants, the acetabular cup is in contact with the ball of the metal or ceramic stem. Every time a movement occurs within the joint, rubbing between the ball and cup occurs and small wear particles of metal and polymer are left behind (see Fig. 10). More often than not, the particles are shed from the softer surface (e.g., ultrahigh molecular weight polyethylene); however, metal particles are also produced. The particles range in size from

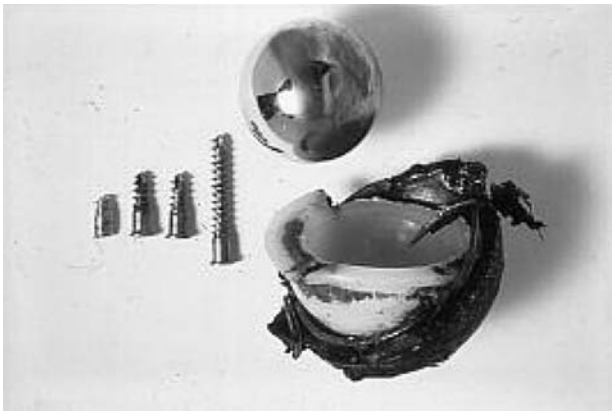


Figure 10. A photograph of some worn biomaterials. Examples in this photograph include screws, a femoral head replacement, and a polyethylene acetabular cup. (Used with permission from the Department of Materials at Queen Mary University of London.)

nanometers to microns with the smaller particles able to enter the lymph fluid and travel to distant parts of the body. The small particles increase the surface area of the material, and this increased surface area can result in increased corrosion (5,13). Thus, wear can lead to deleterious effects (e.g., corrosion), described above, and inflammation, as will be discussed below.

Degradation

Although not affected by corrosion, certain bioactive ceramics and polymers are susceptible to degradation. In the case of bioactive ceramics, however, this process is relatively slow compared with the potential rate of bone regeneration. For polymers, degradation can occur via hydrolytic or, in some cases, enzymatic mechanisms. The chemical structures of both polyamides and polyesters lend themselves toward enzymatic degradation. For polyesters, acidic or alkaline conditions will lead to a deesterification reaction that will eventually destroy the backbone of the polymer. The degradation rate varies greatly depending on the composition of the polymer. For example, within the body, poly(lactic acid) will degrade over many months to years, but poly(glycolic acid) can degrade over a few days or weeks. The degradation rate of polyamides is slower than that of polyesters, but is still an important design consideration when choosing a polymeric biomaterial for a specific application. For applications (e.g., sutures), degradation of the material is a beneficial property since the sutures only need to remain in place for a few days to weeks until the native tissue heals. For applications needing a material with a longer lifespan, degradation poses a larger problem.

Increasingly, degradable polymers or polymers with degradable cross-links are being studied as biomaterials. This interest in degradable systems stems largely from more current research involving tissue engineering and drug delivery (15,16,22,24,26,27). The philosophy of tissue engineering holds that the polymeric biomaterial acts as a scaffold with or without viable cells or biological molecules to promote tissue ingrowth. As cells proliferate and migrate within these scaffolds and begin to create new tissue, the material can and should degrade to leave, ultimately, regenerated or repaired tissue in its place. One of the engineering design constraints, therefore, is to balance the rate of degradation with that of tissue ingrowth. If the biomaterial degrades too rapidly and the newly formed tissue cannot provide the necessary mechanical support, then the biomaterial will have failed. At the same time, if the biomaterial degrades too slowly, then the process of tissue ingrowth may become inhibited or may not occur at all. To this end, more recent research has attempted to include enzymatically sensitive cross-links, usually made from synthetic peptide analogues of enzyme substrates, within polymer networks. Instead of relying upon relatively uncontrolled hydrolytic degradation, the polymeric biomaterial would degrade at a rate controlled by migrating cells. Thus, the cells themselves could degrade the material and produce new tissue in a much more controlled and physiologically relevant manner.

Biological Interactions

Most modern biomaterials are intended to come into direct contact with living tissue and biological fluids. This interaction often makes the biomaterial a target for the protective mechanisms within the body. These protective mechanisms include protein adsorption, hemostasis, inflammation and the foreign body response, and the immune response. Although it has been well established that all types of tissue-contacting biomaterials invoke some degree of biological response, it has only been during the past decade or so when investigations have revealed that all implanted tissue-containing biomaterials invoke an almost identical inflammatory and foreign body response regardless of whether the biomaterial is of metallic, ceramic, polymeric, or composite origin. Although future research in the field of biomaterials aims to better understand and to eventually mitigate the biological interactions that currently result in the failure of many biomaterials, the following biological responses remain of great importance when considering the design and potential applications of any biomaterial. In fact, most current obstacles related to the design of biomaterials involve the interaction of biomaterials with the body and the reaction of the body to biomaterials. As a result, current biomaterial research trends aim to provide an environment that allows the body to invade, remodel, and degrade the implanted material (23,27,28).

Protein Adsorption. As soon as a biomaterial comes into contact with biological fluid (e.g., blood) the material becomes coated with adsorbed proteins. This adsorption is very rapid and is based primarily on noncovalent interactions between various hydrophilic and hydrophobic domains within the adsorbed proteins and the surface of the implanted biomaterial. Initially, the composition of the protein layer depends on the relative concentration of various proteins within the biological fluid. Certain proteins (e.g., albumin) are very abundant in serum and will initially be found abundantly in the adsorbed protein layer. However, over time the adsorbed protein layer will change its composition as proteins with higher affinities for the surface of the material, but lower serum concentrations will displace proteins with lower affinities and higher serum concentrations. This rearrangement and equilibration of the protein layer is known as the Vroman effect. When biomaterials become coated with proteins, surrounding cells no longer see the surface of the material. Instead, they see a layer of serum-soluble proteins. Increasingly, biomaterials design has focused on optimizing surface chemistries and incorporating selective reactive domains that will promote a specific biological response. In reality, these engineered surfaces become masked by a nonspecific protein layer, and it is this protein layer that drives the biological response to an implanted biomaterial. Some successful examples of surface modifications aimed at reducing nonspecific protein adsorption involve the use of nonfouling hydrophilic polymers (e.g., PEG and dextran), the pretreatment of the biomaterial with a specific protein, and the replacement of certain chemically reactive functional groups with others. Time, however, remains the

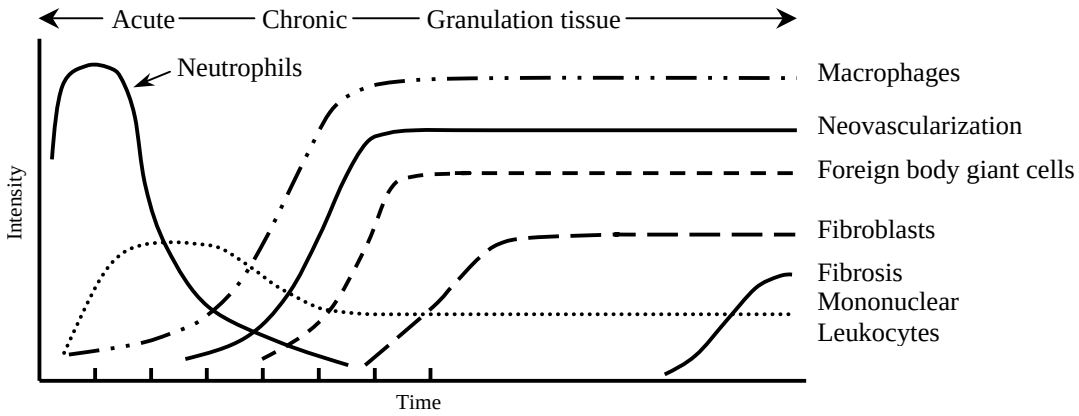


Figure 11. A schematic representing the temporal events involved in acute and chronic inflammation as well as the foreign body response. (Adopted from Anderson et al. as found in Ref. 5.)

largest obstacle with any of these surface treatments. Often, surface treatments will only function for a limited time before serum and other extracellular proteins becomes adsorbed. Once adsorbed, proteins can undergo conformational changes that expose cryptic sites, allow autoactivation, or that influence the behavior of other proteins or cells (5,6,18).

Blood Contact. Direct contact with blood is a major concern of all biomaterials regardless of whether or not they were designed for cardiovascular applications. During surgical implantation, blood vessels are broken, which results in an increased probability that the biomaterial will contact blood. Although exact mechanisms remain unclear, serum proteins (e.g., Factor XII, Factor XI, plasma prekallikrein, and high molecular weight kininogen) interact to initiate contact activation of the coagulation cascade through the intrinsic pathway. This calcium- and platelet-dependent cyclic network involves the activation of thrombin, which ultimately cleaves specific protein domains within fibrinogen and Factor XIII. Activated Factor XIIIa and fibrinogen then react to form a cross-linked fibrin clot. The formation of blood clots as well as the activation of various serum proteins and platelets can lead to local inflammation. Recent approaches have attempted to passivate the blood contact response of implanted biomaterials by incorporating heparin or other antithrombotic agents on the biomaterial surface (5,6,18).

Inflammation and the Foreign Body Response. The human body is well equipped to handle injuries that affect hemostasis. During trauma, proteins within blood can initiate a relatively large biological response lasting days, weeks, and even months. Initially, the area around a trauma site, including the implantation of a biomaterial, becomes inflamed. Inflammation is a normal process involved with healing that is characterized by four major events: swelling, pain, redness, and heat. The vasculature around an injury will become leaky to allow extravasation of various leukocytes (e.g., neutrophils and macrophages). With the presence of cytokines and other growth factors, leukocytes, primarily macrophages, are stimulated to remove bacteria and foreign material. Macrophages also recruit fibroblasts and other cells to the injury site to aid in healing by forming granulation tissue. Over the course of several days or weeks, this initial granulation tissue is remodeled and replaced with restored, functional tissue or, more commonly, scar tissue (Fig. 11).

In the case of implanted biomaterials, the implantation site is the injury site and will become inflamed. As a result, macrophages will be recruited to the site and attempt to remove the "foreign" biomaterial. Unlike smaller injuries, macrophages are unable to remove biomaterials through phagocytosis. When they become frustrated, macrophages will fuse together to form foreign body giant cells. These foreign body giant cells can secrete superoxides and free radicals, which can damage biomaterials, but these cells usually cannot completely remove the foreign biomaterial. In the event that the body cannot eliminate a foreign object through phagocytosis, activated macrophages and foreign body giant cells remain around the implant and can promote a chronic localized area of inflammation. Remaining fibroblasts and other cells around the biomaterial then will begin to secrete a layer of avascular collagen around the biomaterial to effectively encapsulate it and wall it off from the rest of the body (5,6,18). Although the function of some biomaterials is not affected by this foreign body response, biomaterials ranging from sensors to orthopedic implants to soft tissue replacements are adversely affected by this biological reaction. To date, it is not known how to minimize or eliminate an inflammation or foreign body reaction. However, a great deal of research is attempting to create biomaterials that do not evoke a tremendous inflammatory response or that degrade in a way that allows the restoration or repair of native tissue without the adverse affects of chronic inflammation.

Immune Response. The innate and adaptive immune responses of the body also pose a challenge for biomaterials designed for long-term applications. Increasingly, new biomaterials have attempted to incorporate cellular components in an attempt to create new tissues *in vitro* or to seed materials with autologous, allogeneic, or xenogenic cells, including stem cells, to promote tissue repair. Unfortunately, the adaptive immune response will actively eliminate allogeneic or xenogenic cell types. As a result,

biomaterials have been designed to act as barriers that limit lymphocyte activation. Often, cells are encapsulated in microspheres made from various polymers or layers of polymers. For example, pancreatic Islets of Langerhans from animal and human donors have been encapsulated within polymers [e.g., polysulfones, poly(*N*-isopropylacrylamide)] and alginates, to provide an immunoisolated environment that still retains enough permeability to allow for the diffusion of insulin. One of the major complications of this type of biomaterials design is to balance the creation of volume within the microsphere to accommodate enough Islets to allow for sufficient insulin production with the need to provide appropriate diffusion rates so that the cells within the center of the microsphere remain viable. As more polymeric biomaterials incorporate or consist of peptide and protein motifs, there remains a concern as to whether or not these motifs might elicit an adaptive immune response. Even if protein domains derived from human proteins are incorporated into biomaterials, these domains might not be presented the same way to lymphocytes. As a result, the body may start producing antibodies against these domains, which might also lead to certain forms of autoimmune diseases (29).

Although the adaptive immune system is playing an increasingly important role in the rejection of new types of biomaterials, the innate immune system remains a very large threat to the success of a biomaterial. As mentioned above, proteins bind to biomaterials upon implantation. One of the most abundant proteins within the blood is the complement protein C3. Within the blood, C3 can spontaneously hydrolyze to form an active convertase complex, which can cleave C3 into C3a and C3b. Although C3b is rapidly inactivated within the blood, it can remain active if it binds to a surface (e.g., a biomaterial). As a result, the alternative pathway of the complement system can be activated very rapidly leading to formation of membrane-attack complexes but more importantly, the formation of the soluble anaphylotoxins C3a, C4a, and C5a. These anaphylotoxins induce smooth muscle contraction, increase vascular permeability, recruit phagocytic cells, and promote opsonization by phagocytic cells. These phagocytic cells (e.g., macrophages) have receptors recognizing C3b. As a result, macrophages will attempt to engulf the C3b-coated biomaterial. When this fails, the macrophages will form foreign body giant cells, and the body will attempt to encapsulate the biomaterial in a manner similar to that described above for the inflammation and foreign body response (29). Overall, all of the above mentioned biological responses can affect the performance of any biomaterial, and active biomaterials research is striving not only to better understand the mechanisms of inflammation, protein adsorption, hemostasis, and innate and adaptive immune responses, but also to develop strategies to minimize, eliminate, evade, or alter adverse biological responses to materials.

BIOCOMPATIBILITY

Since biomaterials are intended for direct contact with biologically viable tissue, all biomaterials need to possess some degree of biocompatibility. In a manner similar to that of the term biomaterials, the term biocompatibility has experienced many changing definitions over the past several decades. Initially, biocompatibility implied that the biomaterial remained inert to its surroundings in order to refrain from being toxic, carcinogenic, or allergenic. As the definition of biomaterials evolved to include biologically derived materials and molecules, the term biocompatibility needed to encompass these changes. In 1987, David Williams suggested that biocompatibility is "the ability of a material to perform with an appropriate host response in a specific application" (30). Although there does not yet exist a universal consensus with regard to the definition of the term biocompatibility, the definition proposed by Williams provides enough generality to serve as an adequate and accurate description of biocompatibility.

Instead of remaining inert, biomaterials are becoming increasingly reliant on biochemical reactions and physiological processes in order to serve a useful function. In some cases (e.g., in the case of bone plates and artificial joints), biomaterials can remain inert and still provide satisfactory performance. In other instances (e.g., drug delivery vehicles), tissue engineering applications, and *in vivo* organ replacement therapies, biomaterials not only need to actively minimize or adapt to the surrounding biological responses (e.g., inflammation and foreign body responses), but also need to depend on interactions with surrounding tissues and cells in order to provide a useful function (15,22,24–27). In addition, the performance of traditionally inert biomaterials is being enhanced by incorporating chemical or mechanical modifications that interact with biology at the cellular level. For example, the bone-contacting surfaces of metallic femoral stems, for hip replacement, have been modified to contain bioactive ceramic porous networks or hydroxyapatite crystal networks. These ceramic networks allow better osteointegration of the implant with the host tissue and, in some cases, eliminate the need to use bone sealants (e.g., PMMA).

Obviously, if a successful biomaterial needs to show some level of biocompatibility, then there must exist various testing conditions and manufacturing standards to establish safety controls. Organizations [e.g., ASTM International and the International Organization for Standardization (ISO)] do have guidelines and standards for the testing and evaluation of biomaterial biocompatibility. These regulations include tests include the measuring of cytotoxicity, sensitization, skin irritation, intracutaneous reactivity, acute systemic toxicity, genotoxicity, macroscopic and microscopic evaluation of implanted materials and devices, hemocompatibility, subchronic and chronic toxicity, carcinogenicity, the effect of degradation byproducts, and the effect of sterilization (31). For many of these paramenters, the associated standards dictate the size and shape of the material to be tested, appropriate *in vitro* testing procedures and analysis schemes, and relevant testing and evaluation protocols for *in vivo* experimentation. Although standards related to the manufacturing and performance of some biomaterials exist, there remains a lack of uniform biocompatibility testing standards for new classes of biomaterials that rely heavily upon cellular and tissue interactions or that contain biologically active

molecules. New developments in biologically active biomaterials have resulted in not only nonuniform approaches to biocompatibility testing, but also confusions related to the regulatory classification of new types of biomaterials.

FUTURE DIRECTIONS

As more information becomes available regarding biological responses to materials, mechanisms that control embryonic development and early wound healing, and matrix biology, materials will be designed to more adequately address, promote or inhibit biological responses as needed. As a result, the field of biomaterials will not only incorporate principles from materials science and engineering, but also rely increasingly upon design constraints governed by biology (see Fig. 12). Recent trends in biomaterial research show an increased emphasis in designing materials that better match the biological environment with respect to mechanics and biological signals. Materials promote cell attachment using biologically derived signals, degrade through relevant enzymatic degradation and release and store bioactive factors using methods derived from biology. Continued adaptation of materials to more appropriately interact with the living system will result in devices that work with the body to promote tissue regeneration and healing.

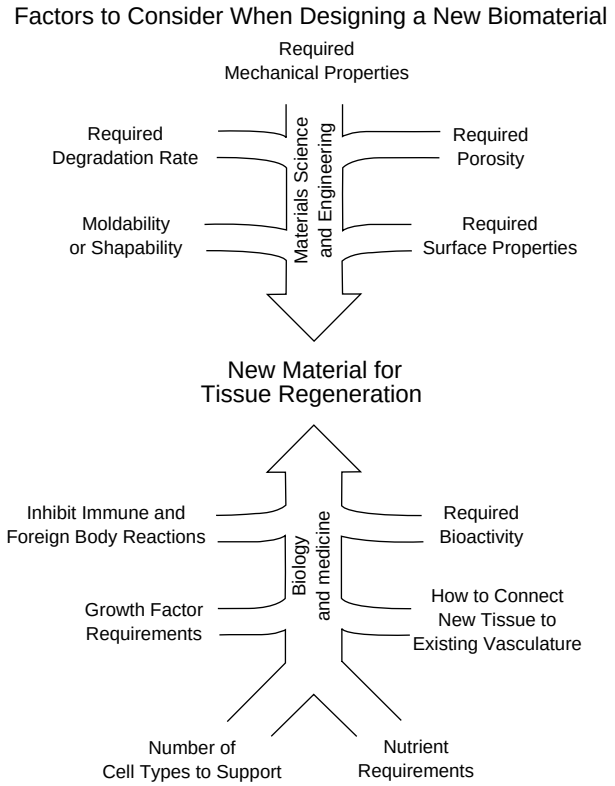


Figure 12. An illustration depicting the various engineering and biological factors that need to be considered in the design of modern biomaterials. Successful, new biomaterials will require the optimization of a variety of parameters and the cooperation of interdisciplinary scientists, engineers, and clinicians. (Reprinted from Ref. 15 with permission from Elsevier.)

BIBLIOGRAPHY

Cited References

1. Black J. The education of the bio-materialist—report of a survey. 1980–1981. J Biomed Mater Res 1982;16(2):159–167.
2. Galletti PM, Boretos JW. Report on the consensus development conference on clinical-applications of biomaterials, 1–3 november 1983. J Biomed Mater Res 1983;17(3):539–555.
3. Lindstrom RL. The polymethylmetacrylate (pmma) intraocular lenses. In: Steinert RF, et al. editors. Cataract surgery: Technique, Complications, and Management. Philadelphia: Saunders; 2003.
4. Barbucci R, editor. Integrated Biomaterials Science. New York: Kluwer Academic; 2002.
5. Ratner BD, Hoffman AS, Schoen FJ, Lemons JE, editors. Biomaterials Science: An Introduction to Materials in Medicine. San Diego: Academic Press; 1996.
6. Greco RS, editor. Implantation Biology: The Host Response and Biomedical Devices. Boca Raton (FL): CRC Press; 1994.
7. Lysaght MJ, O'Loughlin JA. Demographic scope and economic magnitude of contemporary organ replacement therapies. Asaio J 2000;46(5):515–521.
8. Harris G. Step forward for implants with silicone. New York: The New York Times; 2005.
9. Park JB, Lakes RS. Biomaterials: An introduction. 2nd ed. New York: Plenum Press; 1992.
10. Black J, Hastings G, editors. Handbood of Biomaterial Properties. New York: Chapman & Hall; 1998.
11. Wen CE, et al. Novel titanium foam for bone tissue engineering. J Mater Res 2002;17(10):2633–2639.
12. Graham, et al. The use of SMARTeR stents in patients with binary obstrevition. Cliro Radiol 2004;59:288–291.
13. Smith WF, Foundations of Materials Science and Engineering. 2nd ed. New York: McGraw-Hill, Inc.; 1993.
14. Fernandez E, et al. Calcium phosphate bone cements for clinical applications—part ii: Precipitate formation during setting reactions. J Mater Sci-Mater M 1999;10(3):177–183.
15. Seal BL, Otero TC, Panitch A. Polymeric biomaterials for tissue and organ regeneration. Mater Sci Eng R 2001; 34(4–5):147–230.
16. Langer R. 1994 Whitaker lecture—polymers for drug-delivery and tissue engineering. Ann Biomed Eng 1995;23(2): 101–111.
17. Greenwald SE, Berry CL. Improving vascular grafts: The importance of mechanical and haemodynamic properties. J Pathol 2000;190(3):292–299.
18. Dee KC, Puleo DA, Bizios R. An Introduction to Tissue-Biomaterial Interactions. Hoboken (NJ): John Wiley & Sons, Inc.; 2002.
19. Temenoff JS, Mikos AG. Review: Tissue engineering for regeneration of articular cartilage. Biomaterials 2000;21 (5):431–440.
20. Temenoff JS, Mikos AG. Injectable biodegradable materials for orthopedic tissue engineering. Biomaterials 2000;21(23): 2405–2412.
21. Palsson BØ, Bhatia SN. Tissue Engineering. Upper Saddle River (NJ): Pearson Prentice Hall; 2004.
22. Peppas NA, Bures P, Leobandung W, Ichikawa H. Hydrogels in pharmaceutical formulations. Eur J Pharm Biopharm 2000;50(1):27–46.
23. Hubbell JA. Biomaterials in tissue engineering. Bio-Technol 1995;13(6):565–576.
24. Sakiyama-Elbert SE, Hubbell JA. Functional biomaterials: Design of novel biomaterials. Ann Rev Mater Res 2001;31: 183–201.

25. Yamane S, et al. Feasebelety of chitosan-based hyaluoronic acid hybrid biomaterial for a novel scaffold in cartilage tissue engineering. Biomaterials 2005;26:611–619.
26. Gupta P, Vermani K, Garg S. Hydrogels: From controlled release to ph-responsive drug delivery. Drug Discov Today. 2002;7(10):569–579.
27. Ratner BD, Bryant SJ. Biomaterials: Where we have been and where we are going. Annu Rev Biomed Eng 2004;6:41–75.
28. Ratner BD. New ideas in biomaterials science—a path to engineered biomaterials. J Biomed Mater Res 1993;27(7): 837–850.
29. Janeway CA, Travers P, Walport M, Shlomchik MJ. Immunobiology: The Immune System in Health and Disease. 6th ed. New York: Garland Science; 2005.
30. Williams D. Revisiting the definition of biocompatibility. Med Device Technol 2003;14(8):10–13.
31. Bollen LS, Svensen O. Regulatory guidelines for biocompatibility safety testing. Med Plastics Biomater 1997;(May): 16.

See also ALLOYS, SHAPE MEMORY; BIOMATERIALS: BIOCERAMICS; BIOMATERIALS, CORROSION AND WEAR OF; BIOMATERIALS, TESTING AND STRUCTURAL PROPERTIES OF; POLYMERIC MATERIALS.

BIOMATERIALS: BIOCERAMICS

JULIAN R. JONES
LARRY L. HENCH
Imperial College London

INTRODUCTION

During the last century, there has been a revolution in orthopedics with a shift in emphasis from palliative treatment of infection in bone to interventional treatment of chronic age-related ailments. The evolution of stable metallic fixation devices, and the systematic development of reliable total joint prostheses were critical to this revolution in health care. Two alternative pathways of treatment for patients with chronic bone and joint defects are now possible: (*1*) transplantation or (*2*) implantation. Figure 1 shows how approaches to tissue repair have changed and how we think they need to develop.

At present the "gold standard" for the clinical repair of large bone defects is the harvesting of the patient's tissue from a donor site and transplanting it to a host site, often maintaining blood supply. This type of tissue graft (an *autograft*) has limitations; limited availability, morbidity at the donor site, tendency toward resorption, and a compromise in biomechanical properties compared to the host tissue.

A partial solution to some of these limitations is use of transplant tissue from a human donor, a *homograft*, either as a living transplant (heart, heart-lung, kidney, liver, retina) or from cadavers (freeze-dried bone). Availability, the requirement for lifetime use of immunosuppressant drugs, the concern for viral or prion contamination, ethical, and religious concerns all limit the use of homografts.

The first organ transplant (homograft) was carried out in Harvard in 1954. In the United States alone, there are now $>$80,000 organs needed for transplantation at one time, only a quarter of which will be found. The shortage of donors increases every year.

A third option for tissue replacement is provided by transplants (living or nonliving) from other species called *heterografts* or *xenografts*. Nonliving, chemically treated xenografts are routinely used as heart valve replacements (porcine) with $\sim$50% survivability after 10 years. Bovine bone grafts are still in use, but concern of transmission of prions (disease transmission) is growing.

The second line of attack in the revolution to replace tissues was the development of manmade materials to interface with living, host tissues (e.g., implants or prostheses made from biomaterials). There are important advantages of implants over transplants, including

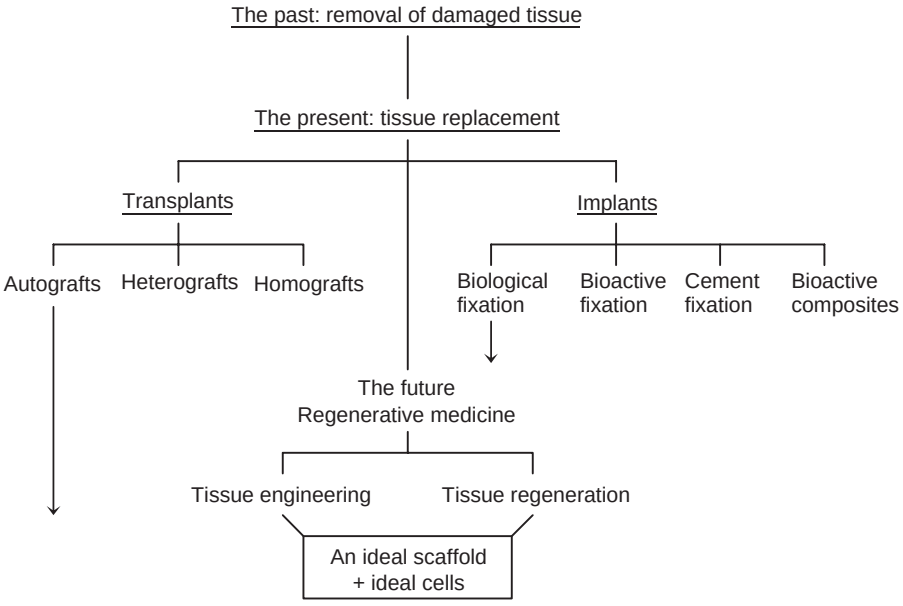


Figure 1. Schematic of the past, present and future of the treatment for diseased and damaged tissue.

availability, reproducibility, and reliability. Failure rate of the materials used in most prostheses are very low, at $< 0.01\%$ (1). As a result, survivability of orthopedic implants such as the Charnley low friction metal–polyethylene total hip replacement is very high up to 15 years (2).

Many implants in use today continue to suffer from problems of interfacial stability with host tissues, biomechanical mismatch of elastic moduli, production of wear debris, and maintenance of a stable blood supply. These problems lead to accelerated wear rates, loosening and fracture of the bone, interface or device that become worse as the patient ages (3). Repair of failed devices, called revision surgery, also becomes more difficult as the patient ages due to decreased quality of bone, reduced mobility, and poorer circulation of blood. In addition, all present day orthopedic implants lack two of the most critical characteristics of living tissues: (*1*) ability to self-repair; and (*2*) ability to modify their structure and properties in response to environmental factors such as mechanical load. The consequences of these limitations are profound. All implants have limited lifetimes. Many years of research and development have led to only marginal improvements in the survivability of orthopedic implants for >15years.

Ideally, artifical implants or devices should be designed to stimulate and guide cells in the body to regenerate tissues and organs to a healthy and natural state. We need to shift our thinking toward regenerative medicine (Fig. 1) (4).

BIOCERAMICS AS MEDICAL DEVICES

A bioceramic is a ceramic that can be implanted into a patient without causing a toxic response. Bioceramics can be classified into three categories; resorbable (e.g., tricalcium phosphate), bioactive (e.g., bioactive glass, hydroxyapatite), and nearly inert materials (e.g., alumina and zirconia) (5). A bioactive material is defined as a material that elicits a specific biological response at the interface of the material, which results in a formation of a bond between the tissue and that material (6). Bioceramics can be polycrystalline (alumina or hydroxyapatite), bioactive glass, bioactive glass–ceramic (apatite/ wollastonite, A/W), or used in bioactive composites such as polyethylene–hydroxyapatite.

This article begins with examples of successful bioceramics used clinically that improve the length and quality of life for patients. Developments of porous bioceramic and composite scaffolds for tissue engineering applications are then discussed. The article ends by discussing how bioactive ceramics may be the future of regenerative medicine due to their potential for guiding tissue regeneration by stimulating cells at the genetic level (Fig. 1).

NEARLY INERT BIOCERAMICS

High density, high purity α-alumina (Al_2O_3) was the first bioceramic widely used clinically, as the articulating surfaces of the ball and socket joints of total hip replacements because of its combination of low friction, high wear resistance, excellent corrosion resistance, good biocompatibility, and high strength (7). The physical properties of alumina depend on the grain size. Medical grade alumina exhibits an average grain size <4 μm, a compressive strength of 4.5 GPa and a Young's modulus of 400 GPa (8). Other clinical applications of Al_2O_3 include bone screws, jaw and maxillofacial reconstructions, middle ear bone substitutions, and dental implants.

Zirconia is a bioinert ceramic that has higher flexural strength and fracture toughness and a lower Young's modulus than alumina. Zirconia may therefore be suitable for bearing surfaces in total hip prostheses, however, there are concerns over the wear rate and radioactivity of the material in the body (9).

When an almost inert implant is implanted into soft or hard tissue, a nonadherent fibrous capsule surrounds the implant. If the implant is loaded such that interfacial movement can occur, the fibrous capsule can become several hundred micrometers thick and cause loosening of the implant, which will eventually lead to clinical failure (8).

An improved interface between nearly inert implants and tissue can be achieved by using an implant containing pores in excess of 100 μm in diameter. The fibrous connective (scar) tissue grows into these pores and anchors the implant in place. This technique is termed "biological fixation" (10). Viable bone requires pores of >200 μm. However, connective tissue still allows some movement of the prothesis, which will increase with age and cause bone resorption.

THE CHALLENGE FOR BIOCERAMICS

Bone is a natural composite of collagen (type I) fibers, noncollagenous proteins, and mineralized bone. It is a rigid material that exhibits a hierarchical structure with an outer layer of dense cortical bone and an internal structure of porous cancellous and trabecular (spongy) bone. Trabecular bone is orientated spongy bone that is found at the end of long bones and in vertebras (11). This structure provides excellent mechanical properties: cortical bone exhibits a compressive strength of 100–230 MPa and a Young's modulus of 7–30 GPa; cancellous bone exhibits a compressive strength of 2–12 MPa and a Young's modulus of 0.05–0.5 GPa (8). Bone is generated by cells called osteoblasts and resorbed by cells called osteoclasts, which remodel the bone in response to external stimuli such as mechanical load (11). In order to regenerate bone, the implant should exhibit a Young's modulus similar to that of the bone. If the modulus of the implant is higher than the bone then stress shielding can occur, where the stem supports the total load. If this occurs, osteoclasts resorb bone from the implant interface (12). An example of this is the use of alumina as in total hip replacements. Alumina exhibits a modulus 10–50 times higher than cortical bone. If the modulus of the implant is substantially lower than the bone, the implant is unlikely to be able to withstand the loading environment and will fracture.

Ceramics have the potential to prevent stress-shielding and have many properties that can aid bone regeneration (8). Therefore, we will concentrate on the use of bioceramics

in orthopedics, but will also describe adaption for soft tissue applications.

Osteoporosis is a condition where the density and strength of the trabecular bone decreases (13), due to osteoblasts becoming progressively less active and the pore walls (trabeculae) in the internal spongy bone are reduced in thickness and number causing spinal problems, hip fracture, and subsequent hip replacement operations.

The challenge for bioceramics is to replace old, deteriorating bone with a material that can function for as long as is needed, which may be > 20 years. There are two options to satisfy increasing needs for orthopedic repair in the new millennium: (*1*) improve implant survivability by 10–20 years; or (*2*) develop alternative devices that can regenerate tissues to their natural form and function. Decades of research have not been able to achieve the first, discussion of the second, the application of bioactive bioceramics, and their role in regenerative medicine, particularly in bone regeneration follows.

RESORBABLE BIOCERAMICS

Tricalcium phosphate (TCP) resorbs on contact with body fluid. Resorbable materials are designed to dissolve at the same rate that a tissue grows, so that they eventually are totally replaced by the natural host tissue. However, matching the resorption rate of TCP with bone growth is difficult and since TCP ceramics exhibit low mechanical strength, so they cannot be used in load bearing applications (14).

THE BIOACTIVE ALTERNATIVE

During the last decade, considerable attention has been directed toward the use of bioceramic implants with bioactive fixation, where bioactive fixation is defined as interfacial bonding of an implant to tissue by means of formation of a biologically active hydroxyapatite layer on the surface of the implant. This layer bonds to the biological apatite in bone (8).

An important advantage of bioactive fixation is that a bioactive bond forms at the implant–bone interface with a strength equal to, or greater than, bone after 3–6 months. The level of bioactivity of a specific material can be related to the time taken for $> 50\%$ of the interface to bond to bone ($t_{0.5bb}$) (15);

$$\text{Bioactivity index}, I_B = 100/t_{0.5bb} \quad (1)$$

Materials exhibiting an I_B value > 8 (class A), will bond to both soft and hard tissue. Materials with an I_B value < 8 (class B), but > 0, will bond only to bone. Biological fixation is capable of withstanding more-complex stress states than implants that only achieve morphological fixation, that is surface fixation to roughness (15). There are a number of bioactive bioceramics.

SYNTHETIC HYDROXYAPATITE

Synthetic hydroxyapatite (HA) $Ca_{10}(PO_4)_6(OH)_2$ has been developed to match the biological apatite in bone. Biological apatite, although similar, exhibits different stoichiometry, composition, and crystallinity to pure HA. Biological apatites are usually calcium deficient and carbonate substituted (primarily for phosphate groups) (16). Hydroxyapatite is a class B bioactive material, that is, it bonds only to bone and promotes bone growth along its surface (osteoconduction). The mechanism for bone bonding involves the development of a cellular bone matrix on the surface of the HA, producing an electron dense band 3–5 μm wide. Collagen bundles appear between this area and the cells. On contact with body fluid, a dissolution–precipitation process occurs at the HA surface resulting in the formation of carbonated apatite microcrystals, which are similar to biological HA and are incorporated into the collagen. As the site matures, collagen fibrils mineralize and the interfacial layer decreases in thickness as crystallites of the growing bone align with those of the implant. Commercial production methods for synthetic HA usually involve a dropwise addition of phosphoric acid to a stirring suspension of calcium hydroxide in water, which causes an apatite precipitate to form. Ammonia is added to keep the pH very alkaline and to ensure formation of HA when the precipitate is sintered at 1250 °C. Commercial dense HA exhibits a compressive strength in excess of 400 MPa and a Young's modulus of 12 GPa. There are many clinical applications for HA implants including the repair of bony defects and tooth root replacement (16).

Hydroxyapatite has also been used as a plasma-sprayed coating on porous metallic implants in total hip replacements, allowing a bond to form between the bone and the implant (17). Initial bone ingrowth is more rapid than uncoated porous metallic implants, but the long-term survivability of the implants will not be known until after 10-year follow-up clinical trails have been completed.

BIOACTIVE GLASSES

Bioactive glasses are class A bioactive materials (I_B value > 8) that bond to soft and hard tissue and are osteoconductive, which means that bioactive glass implants stimulate bone formation on the implant away from the host bone–implant interface (15). Bioactive glasses undergo surface dissolution in a physiological environment in order to form a hydroxycarbonate apatite (HCA) layer. This is very similar to the carbonate-substituted apatite in bone. The higher the solubility of a bioactive glass, the more pronounced is the effect of bone tissue growth. The structures of bioactive glasses are based on a cross-linked silica network modified by cations. The original bioactive glasses were developed by Hench and colleagues in the early 1970s (18) and were produced using conventional glass melt-processing techniques with a composition of 45S5, 46.1% SiO_2, 24.4% NaO, 26.9% CaO, and 2.6% P_2O_5, in mol percent. This composition was given the name Bioglass.

MECHANISM OF BIOACTIVITY OF BIOACTIVE GLASSES

When a glass reacts with an aqueous solution, both chemical and structural kinetic changes occur as a function of time within the glass surface (8). Accumulation of

dissolution products causes both the chemical composition and pH of solution to change. The formation of HCA on bioactive glasses and the release of soluble silica to the surrounding tissue are key factors in the rapid bonding of these glasses to tissue and the stimulation of tissue growth.

There are 11 stages in process of complete bonding of bioactive glass to bone. Stages 1–5 are chemical; stages 6–11 are biological (4,15);

1. Rapid exchange of Na^+ and Ca^{2+} with H^+ or H_3O^+ from solution (diffusion controlled with a $t^{1/2}$ dependence, causing hydrolysis of the silica groups, which creates silanols;

$$Si-O-Na^+ + H^+ + OH^- \rightarrow Si-OH^+ + Na^+_{(aq)} + OH^-$$

   The pH of the solution increases as a result of H^+ ions in the solution being replaced by cations.
2. The cation exchange increases the hydoxyl concentration of the solution, which leads to attack of the silica glass network. Soluble silica is lost in the form of $Si(OH)_4$ to the solution, resulting from the breaking of Si–O–Si bonds and the continued formation of Si–OH (silanols) at the glass solution interface:

$$Si-O-Si + H_2O \rightarrow Si-OH + OH-Si$$

   This stage is an interface-controlled reaction with a $t^{1.0}$ dependance.
3. Condensation and repolymerization of a SiO_2-rich layer on the surface, depleted in alkalis and alkali earth cations:
4. Migration of Ca^{2+} and $PO_4{}^{3-}$ groups to the surface through the SiO_2-rich layer, forming a $CaO–P_2O_5$-rich film on top of the SiO_2-rich layer, followed by growth of the amorphous $CaO–P_2O_5$-rich film by incorporation of soluble calcium and phosphates from solution.
5. Crystallization of the amorphous $CaO–P_2O_5$ film by incorporation of OH^- and $CO_3{}^{2-}$ anions from solution to form a mixed-HCA layer.
6. Adsorption and desorption of biological growth factors, in the HCA layer (continues throughout the process) to activate differentiation of stem cells.
7. Action of macrophages to remove debris from the site.
8. Attachment of stem cells on the bioactive surface.
9. Differentiation of stem cells to form bone growing cells, such as osteoblasts.
10. Generation of extra cellular matrix by the osteoblasts to form bone.
11. Crystallization of inorganic calcium phosphate matrix to enclose bone cells in a living composite structure.

Interfacial bonding occurs with bone because of the biological equivalence of the inorganic portion of bone and the growing HCA layer on the bioactive implant. For soft tissues, the collagen fibrils are chemisorbed on the porous SiO_2-rich layer via electrostatic, ionic and/or hydrogen bonding, and HCA is precipitated and crystallized on the collagen fiber and glass surfaces.

Reaction stages one and two are responsible for the dissolution of a bioactive glass, and therefore greatly influence the rate of HCA formation. Studies have shown that the leaching of silicon and sodium to solution, from melt-derived bioactive glasses, is initially rapid, following a parabolic relationship with time for the first 6 h of reaction, then stabilizes, following a linear dependence on time, which agree with the dissolution kinetics of soda lime–silica glasses:

$$Q = Kt^{\gamma} \quad \text{for total diffusion, or more generally} \tag{2}$$

$$Q = at^{\gamma} + bt \quad \text{for total diffusion and selective leaching} \tag{3}$$

where Q is the quantity of alkali ions from the glass, t is the duration of experiment, a,b are empirically determined constants, K is the reaction rate constant, assuming constant glass area and temperature, and $\gamma = 1/2$ (for stage 1) or 1 (for stage 2); as $t \rightarrow 0$ $\gamma = 1/2$, as $t \rightarrow \infty$ $\gamma = 1$ (19).

Phosphorous and calcium contents of the solution follow a similar parabolic trend over the first few hours, after which they decrease, corresponding with the formation of the Ca–P-rich film (stage 4). The pH change of the solution mirrors dissolution rates. An initial rapid increase of pH is a result of ion exchange of cations such as Na^+ from the glass with H^+ from solution during the first minutes of reaction at the bioactive glass surface. As release rate of cations decreases, the solution pH value tends toward a constant value.

For bioactive implants, it is necessary to control the solubility (dissolution rate) of the material. A low solubility material is needed if the implant is designed to have a long life, for example, a coating on orthopedic metals, such as synthetic hydroxyapatite on a titanium alloy femoral stem. A high solubility implant is required if it is designed to aid bone formation, such as 45S5 Bioglass powders for bone graft augmentation. A fundamental understanding of factors influencing solubility and bioreactivity is required when developing new materials for *in situ* tissue regeneration and tissue engineering.

FACTORS AFFECTING THE DISSOLUTION AND BIOACTIVITY OF GLASSES

Many factors affect the dissolution rate, and therefore bioactivity of bioactive glasses. The composition, initial pH, ionic concentration, and temperature of the aqueous environment have a large effect on the dissolution of the glass. The presence of proteins in the solution has been found to reduce dissolution rates due to the adsorption of serum proteins onto the surface of the bioactive glass, which form a barrier to nucleation of the HCA layer (19).

A change in geometry and surface texture of an implant will generally mean a change in the surface area/solution volume ratio (SA/V). An increase in the SA/V generally causes an increase in the dissolution rate, as the amount of surface exposed to solution for ion exchange increases. An increase in SA/V can be caused by a decrease in particle size or by an increase in surface roughness or porosity (19). A

similar effect occurs if the volume of surrounding solution increases (20).

If silicate glasses are considered to be inorganic polymers of silicon cross-linked by oxygen, the network connectivity is defined as the average number of cross-linking bonds for elements other than oxygen that form the backbone of a silicate glass network. The network connectivity can be used to predict solubility (21). Calculation of network connectivity is based on the relative numbers of network forming oxide species (those that contribute to cross-linking or bridging) and network-modifiers (nonbridging) present. Silicate structural units in a glass of low network connectivity are probably of low molecular mass and capable of going into solution. Consequently, glass solubility increases as network connectivity is reduced. The network connectivity can be used to predict bioactivity. Crystallization of a glass inhibits its bioactive properties, because ion exchange is inhibited by crystalline phases, and interferes with network connectivity.

Slight deviations in glass composition can radically alter the dissolution kinetics and the basic mechanisms of bonding. It is widely accepted that increasing silica content of melt-derived glass decreases dissolution rates by reducing the availability of modifier ions such as Ca^{2+} and HPO^{4-} to the solution and the inhibiting development of a silica-gel layer on the surface. The result is the reduction and eventual elimination of the bioactivity of the melt-derived bioactive glasses as the silica content approaches 60%. The addition of multivalent cations, such as alumina, stabilizes the glass structure by eliminating nonbridging oxygen bonds reducing the rate of break-up of the silica network and reducing the rate of HCA formation. Melt-derived glasses with >60 mol% silica are not bioactive. In order to obtain bioactivity at silica levels >60 mol%, the sol–gel process is used, which is a novel processing technique for the synthesis of tertiary bioactive glasses.

CLINICAL APPLICATIONS OF MELT-DERIVED BIOACTIVE GLASSES

Bioactive glasses have been used for >15 years to replace the small bones of the middle ear (ossicles) damaged by chronic infection (22). The glass bonds to the soft tissue of the eardrum and to the remaining bone of the stapes footplate, anchoring both ends of the implant without the formation of fibrous (scar) tissue.

In 1993, particulate bioactive glass, 45S5 Bioglass was cleared in the United States for clinical use as a bone graft material for the repair of periodontal osseous defects (Perioglas, USBiomaterials Alachua, Florida). The glass powder is inserted into the cavities in the bone between the tooth and the periodontal membrane and the tooth, which have eroded due to periodontitis. New bone is rapidly formed around the particles restoring the anchorage of the tooth in place (23). Since 1993, numerous oral and maxillofacial clinical studies have been conducted to expand the use of this material. More than 2,000,000 reconstructive surgeries in the jaw have been done using Perioglas. The same material has been used by several orthopedic surgeons to fill a variety of osseous defects and for clinical use in orthopedics, now termed NovaBone, it is now approved for clinical use worldwide.

GLASS-CERAMICS

Bioactive glasses and sintered HA do not have mechanical properties as high as that of cortical bone. Kokubo et al. (24) developed dense apatite/ wollastonite (A/W) glass-ceramics by heating crushed quenched melt-derived glass (MgO 4.6, CaO 44.7, SiO_2 34.0, P_2O_5 6.2, CaF_2 0.5 wt%) to 1050 °C at a rate of 5 °C · min. Oxyapatite (38 wt%) and β-wollastonite (34 wt%) precipitated and were homogeneously dispersed in a glassy matrix (28 wt%). A/W glass-ceramic (Cerabone) has a compressive strength of 1080 MPa and a Young's modulus of 118 GPa, an order of magnitude higher than cortical bone. On contact with body fluid, the A/W glass-ceramic forms a surface layer of carbonated apatite (HCA) similar to biological apatite. The release of calcium to solution causes a hydrated silica layer to form on the glass phase, providing nucleation sites for the HCA layer. Figure 2 shows how A/W glass-ceramics bond to bone more rapidly than sintered HA, but less rapidly than Bioglass. The A/W glass-ceramics are not resorbable, but due to the high compressive strengths A/W glass-ceramics are used as replacement vertebrae, iliac prostheses and in a granular form as bone defect fillers.

SOL–GEL-DERIVED BIOACTIVE GLASSES

Until the late 1980s, bioactive glasses were generally melt-derived, with the majority of research aimed at the 45S5 Bioglass composition (46.1% SiO_2, 24.4% NaO, 26.9% CaO,

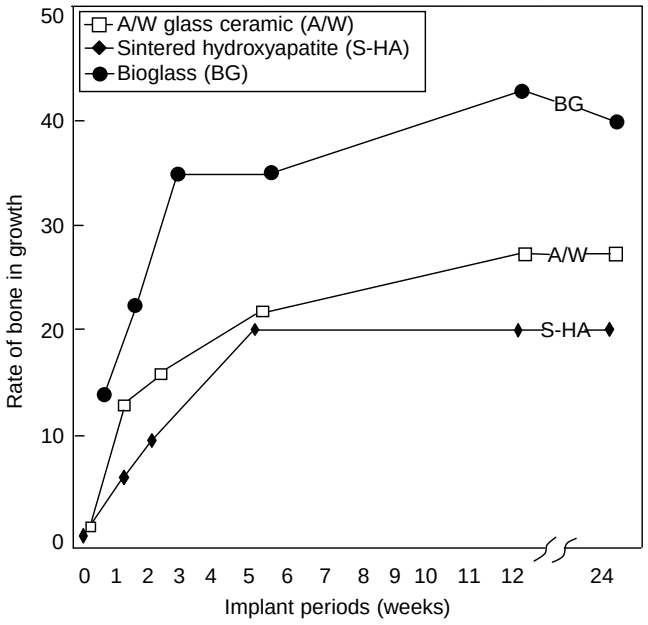


Figure 2. Graph of rate of bone ingrowth into the spaces between bioactive ceramic particles (diameters 300–500 μm) as a function of implantation time for Bioglass (BG), A/ W glass–ceramics (A/ W) and sintered hydroxyapatite (s-HA).

and 2.6% P_2O_5, in mol%) and apatite-wollastanite (A/W) glass-ceramics. The recognition that the silica gel layer plays a role in HCA nucleation and crystallization led to the development of the bioactive three component $CaO–P_2O_5–SiO_2$ sol–gel-derived glasses by Li et al. (25).

A sol is a dispersion of colloidal particles (solid particles with diameters 1–100 nm) in a liquid. A gel is an interconnected, rigid network of polymeric chains with average lengths > 1 μm in a continuous fluid phase and pores of submicrometer dimensions.

There are three methods that are used to produce sol–gel monoliths (26):

1. Network growth of discrete colloidal powders in solution.
2. Simultaneous hydrolysis and polycondensation of alkoxide or nitrate precursors, followed by hypercritical point drying of gels.
3. Simultaneous hydrolysis and polycondensation of alkoxide precursors, followed by ageing and drying under ambient conditions.

The pore liquid is removed from the three-dimensional (3D) network as a gas phase. A gel is defined as dried when the physically absorbed water is completely evacuated. This occurs between 100 and 180 °C. Homogeneous gel–glasses are only obtained when a sol–gel method using alkoxide precursors is employed (i.e., methods 2 or 3).

Liquid alkoxide precursors, such as $Si(OR)_4$, where R is CH_3, C_2H_5, and so on, are mixed with a solvent (usually water) and a catalyst. Tetraethylorthosilicate (TEOS) and tetramethylorthosilicate (TMOS) are the alkoxide precursors most commonly used for sol–gel derived silica. Hydrolysis and condensation (aqueous and alcoholic) reactions follow, forming a 3D SiO_2 network of continuous Si–O–Si links that span throughout the solvent medium. When sufficient interconnected Si–O–Si bonds are formed in a region, they respond cooperatively as a sol. A silica network is formed by the condensation reactions and the sol becomes a gel when it becomes rigid and it can support a stress elastically. The gel point is characterized by a steep increase in elastic and viscous moduli. A highly interconnected 3D gel network is obtained, composed of $(SiO_4)_4$ tetrahedra bonded either to neighboring silica tetrahedra via bridging oxygen (BO) bonds or by Si–O–Ca or Si–O–P nonbridging (NBO) bonds. The gel consists of interpenetrating solid and liquid phases: the liquid (a byproduct of the polycondensation reactions) prevents the solid network from collapsing and the solid network prevents the liquid from escaping. The gel is aged at ~60 °C to allow cross-linking of silica species and further network formation. The liquid is then removed from the interconnected pore network by evaporation of the solvent at elevated temperature to form a "xerogel". During this stage, the gel network undergoes considerable shrinkage and weight loss. This stage is critical in obtaining crack-free bodies as large capillary stresses can develop due to solvent evaporation through the pore network. To prevent cracking, the drying process must be controlled by decreasing liquid surface energy, by controlling the rates of hydrolysis and condensation using the precursors, and controlling the thermal drying conditions carefully. Extending the ageing time can help prevent cracking during drying (27). However, even under optimum conditions it is difficult to produce multicomponent crack-free silica-based glasses with diameters in excess of 10 mm.

Dried xerogels have a very large concentration of silanols on the surface of the pores, which renders them chemically unstable at room temperature. The gel is stabilized by sintering at > 500 °C, which removes chemically active surface groups (such as silanols or trisiloxane rings) from the pore network, so that the surface does not rehydroxylate in use. Thermal methods are most common, but chemical methods, involving replacing silanols with more hydrophobic and less reactive species are also possible.

In multicomponent systems, the stabilization process also decomposes other species in the gel such a nitrates or organics. In this thesis nitrates are present after drying. Such species are a source of inhomogeneity and are biologically toxic. Pure $Ca(NO_3)_2$ decomposes at 561 °C, therefore thermal stabilization must be carried out above this temperature.

Sol–gel derived bioactive glasses exhibit a mesoporous texture, that is, pores with diameters in the range 2–50 nm that are inherent to the sol–gel process. The textural properties of the glass are affected by each stage of the sol–gel process, that is, temperature, sol composition, ageing, drying rate, and stabilization temperatures and rates.

Advantages of Sol–Gel-Derived Glasses

There are several advantages of a sol–gel-derived glass over a melt-derived glass, which are important for biomedical applications. Sol–gel-derived glasses have (26):

1. Lower processing temperatures (600–700 °C for gel–glasses compared to 1100–1300 °C for melt-derived glasses).
2. The potential of improved purity, required for optimal bioactivity due to low processing temperatures and high silica and low alkali content.
3. Improved homogeneity.
4. Wider compositions can be used (up to 90 mol% SiO_2) while maintaining bioactivity.
5. Better control of bioactivity by changing composition or microstructure.
6. Structural variation can be produced without compositional changes by control of hydrolysis and polycondensation reactions during synthesis.
7. A greater ease of powder production.
8. Interconnected nanometer scale porosity that can be varied to control dissolution kinetics or be impregnated with biologically active phases such as growth factors.
9. A higher bioactivity due to the textural porosity (SA/V ratio two orders of magnitude higher than melt-derived glasses).
10. Gel–glasses are resorbable and the resorption rate can be controlled by controlling the mesoporosity.

11. Can be foamed to provide interconnected pores of 10–200 μm, mimicking the architecture of trabecular bone.

The mechanism for HCA formation on bioactive gel–glasses follows most of the same 11 stages as those for melt-derived glasses except that dissolution rates are much higher due to the mesoporous texture which creates a higher SA/ V ratio, increasing the area of surface exposed for cation exchange (stage 1) and silica network break-up (stage 2). There are also more sites available for HCA layer formation (19).

FEATURES OF CLASS A BIOACTIVE MATERIALS

An important feature of Class A bioactive materials is that they are osteoproductive as well as osteoconductive. In contrast, Class B bioactive materials exhibit only *osteoconductivity*, defined as the characteristic of bone growth and bonding along a surface. Dense synthetic HA ceramic implants exhibit Class B bioactivity. *Osteoproduction* occurs when bone proliferates on the surfaces of a material due to enhanced osteoblast activity. Enhanced proliferation *and* differentiation of osteoprogenitor cells, stimulated by slow resorption of the Class A bioactive particles, are responsible for osteoproduction.

Is Bioactive Fixation the Solution?

During the last decade, it has been assumed that improved interfacial stability achieved with bioactive fixation would improve implant survivability. Clinical trials have shown this to often not be the case. Replacement of the roots of extracted teeth with dense HA ceramic cones to preserve the edentulous alveolar ridge of denture wearers resulted in generally $<50\%$ survived at only 5 years. Early use of HA-coated orthopedic implants seldom survived 10 years $>85\%$ figure for cemented total hip prostheses (1). However, long-term success rates of bioactive HA coatings have improved during the last decade due to greater control of the coating process. The survivability of HA coated femoral stems is now equivalent at 10 years to cemented prostheses. It will take another 5 years to know if survivability is superior when HA coatings are used.

Why is bioactive fixation not a panacea to hip implant survivability? There are three primary reasons: (*1*) metallic prostheses with a bioactive coating still have a mismatch in mechanical properties with host bone, and therefore less than optimal biomechanical and bioelectric stimuli, at the bonded interface; (*2*) the bioactive bonded interface is unable to remodel in response to applied load; and (*3*) use of bioactive materials does not solve the problem of osteolysis due to wear debris generated from the polyethylene cups. Use of alumina–alumina bearing surfaces eliminates most wear debris from total hip prostheses, but increases the cost of the prosthesis by 200–300%. For younger patients the cost is acceptable, but for the general population it often is considered to be too expensive.

Most biomaterials in use today and the prostheses made from the materials have evolved from trial and error experiments. Optimal biochemical and biomechanical features that match living tissues have not been achieved, so it also is not surprising that long-term implant survivability has not been improved very much during the last 15 years.

THE BIOCOMPOSITES ALTERNATIVE

Bone is a natural composite of collagen fibers (polymer) and mineral (ceramic). Therefore to create an implant that mimics the mechanical properties of bone, a composite should provide high toughness, tensile strength, fatigue resistance, and flexibility while maintaining modulus similar to bone. Biocomposites are being developed to eliminate elastic modulus mismatch and stress shielding of bone. Two approaches have been tried. Bioinert composites, such as carbon–carbon fiber composite materials, are routinely used in aerospace and automotive applications. These lightweight, strong, and low modulus materials would seem to offer great potential for load-bearing orthopedic devices. However, delamination can occur under cyclic loading that releases carbon fibers into the interfacial tissues. The carbon fibers can give rise to a chronic inflammatory response. Thus, bioinert composites are not widely used and are unlikely to be a fruitful direction for development in the next decade.

BIOACTIVE COMPOSITES

The second approach is to make a bioactive composite that does not degrade, such as pioneered by at the IRC in Biomedical Materials, University of London. Bonfield and co-workers (28) increased the stiffness of a biocompatible polymer (polyethylene) from 1 to 8 GPa by adding a secondary phase with higher modulus (HA). The compressive strength of the composite, now called HAPEX, was 26 MPa. Addition of HA also meant that the composite would also bond to bone. Applications for HAPEX have included ossicular replacement prostheses and the repair of orbital floors in the eye socket. Ideally, it is possible to match the properties of both cancellous and cortical bone, although this is seldom achieved by the biocomposites available today. A challenge for the next decade is to use advanced materials processing technology to improve the interfacial bonding between the phases and reduce the size of the second-phase particles, thereby increasing the strength and fracture toughness of these new materials.

Another option is to use a resorbable polymer matrix for a biocomposite that will be replaced with mineralizing bone as the load on the device is increased. Work in this area is in progress, but it is difficult to maintain structural integrity as resorption occurs. The tissue engineering alternative is based upon this concept (29). Further details on biomedical composites can be found in a review by Thompson and Hench (30).

A NEW REVOLUTION IN ORTHOPEDICS?

We suggest that the orthopedics revolution of the last 30 years, the revolution of replacement of tissues by transplants and implants, has run its course. It has led to a

remarkable increase in the quality of life for millions of patients; total joint prostheses provide excellent performance and survivability for 15–20 years. Prostheses will still be the treatment of choice for many years to come for patients of 70 years or older. However, continuing the same approach of the last century; that is, modification of implant materials and designs is not likely to reach a goal of 25–30 years implant survivability, an increasing need of our ageing population. We need a change in emphasis in orthopedic materials research; in fact, we need a new revolution.

BIOCERAMICS IN REGENERATIVE MEDICINE

The challenge for the next millennium in bioceramics and biomedical materials in general is to shift the emphasis of research toward assisting or enhancing the body's own reparative capacity. We must recognize that within our cells lies the genetic information needed to replicate or repair any tissue. We need to learn how to activate the genes to initiate repair at the right site.

Our goal of regeneration of tissues should involve the restoration of metabolic and biochemical behavior at the defect site, which would lead to restoration of biomechanical performance, by means of restoration of the tissue structure leading to restoration of physiological function.

The concept requires that we develop biomaterials that behave in a manner equivalent to an autograft, that is, what we seek is a *regenerative allograft* or *scaffold*. This is a great challenge. However, the time is ripe for such a revolution in thinking and priorities. Regenerative medicine encompasses many fields. We concentrate here on the use of bioceramics in tissue engineering and regeneration applications that require scaffolds to promote tissue repair. Tissue regeneration techniques involve the use of a scaffold that can be implanted into a defect to guide and stimulate tissue regrowth *in situ*. The scaffold should resorb as the tissue grows, leaving no trace. In tissue engineering applications, the scaffolds are seeded with cells *in vitro* to produce the basis of a tissue before implantation; cells extracted from a patient, seeded on a scaffold of the desired architecture and the replacement tissue grown in the laboratory, ready for implantation. The use of the patient's own cells from the same patient would eliminate any chance of immunorejection (31).

GENETIC CONTROL BY BIOACTIVE MATERIALS

We have now discovered the genes involved in phenotype expression and bone and joint morphogenesis, and thus are on the way toward learning the correct combination of extracellular and intracellular chemical concentration gradients, cellular attachment complexes, and other stimuli required to activate tissue regeneration *in situ*. Professor Julia Polak's group at the Imperial College London Centre for Tissue Engineering and Regenerative Medicine has recently shown that seven families of genes are up- and down-regulated by bioactive glass extracts during proliferation and differentiation of primary human osteoblasts *in vitro* (32). These findings should make it possible to design a new generation of bioactive materials for regeneration of bone. The significant new finding is that low levels of dissolution of the bioactive glass particles in the physiological environment exert a genetic control over osteoblast cell cycle and rapid expression of genes that regulate osteogenesis and the production of growth factors.

Xynos et al. (33) showed that within 48 h a group of genes was activated including genes encoding nuclear transcription factors and potent growth factors. These results were obtained using cultures of human osteoblasts, obtained from excised femoral heads of patients (50–70 years) undergoing total hip arthroplasy.

In particular, insulin-like growth factor (IGF) II, IGF-binding proteins, and proteases that cleave IGF-II from their binding proteins were identified (34). The activation of numerous early response genes and synthesis of growth factors was shown to modulate the cell cycle response of osteoblasts to the bioactive glasses and their ionic dissolution products. These results indicate that bioactive glasses enchance osteogenesis through a direct control over genes that regulate cell cycle induction and progression. However, these molecular biological results also confirm that the osteoprogenitor cells must be in a chemical environment suitable for passing checkpoints in the cell cycle toward the synthesis and mitosis phases. Only a select number of cells from a population are capable of dividing and becoming mature osteoblasts. The others are switched into apoptosis and cell death. The number of progenitor cells capable of being stimulated by a bioactive medium decreases as a patient ages, which may account for the time delay in formation of new bone in augmented sites.

Enormous advances have been made in developmental biology, genetic engineering, cellular and tissue engineering, imaging and diagnosis, and in microoptical and micromechanical surgery and repair. Few of these advances have, as yet, been incorporated with the molecular design of new biomaterials. This must be a high priority for the next two decades of research. However, for large defects a scaffold is required to guide tissue regeneration in 3D. Ideally, the scaffold should also release active agents that can also stimulate the cells within the tissue.

AN IDEAL SCAFFOLD

An ideal scaffold is one that mimics the extracellular matrix of the tissue that is to be replaced so that it can act as a 3D template on which cells attach, multiply, migrate, and function. The criteria for an ideal scaffold for bone regeneration are that it (35,36):

1. Is made from a material that is biocompatible (i.e., not cytotoxic).
2. Acts as template for tissue growth in 3D.
3. Has an interconnected macroporous network containing pores with diameters in excess of 100 μm for cell penetration, tissue ingrowth and vascularization, and nutrient delivery to the center of the regenerating tissue on implantation.
4. Bonds to the host tissue without the formation of scar tissue (i.e., is made from an bioactive and osteoconductive–osteoproductive material).

5. Exhibits a surface texture that promotes cell adhesion, adsorption of biological metabolites.
6. Influences the genes in the bone generating cells to enable efficient cell differentiation and proliferation.
7. Resorbs at the same rate as the tissue is regenerated, with degradation products that are nontoxic and that can be easily be excreted by the body, for example, via the respiratory or urinary systems. Is made from a processing technique that can produce irregular shapes to match that of the defect in the patient. Has the potential to be commercially producible to the required ISO (International Standards Organization) or FDA (Food and Drug Administration) standards.
8. Can be sterilized and maintained as a sterile product to the patient.
9. Can be produced economically to be covered by national and/ or private healthcare insurances.

For *in situ* bone regeneration applications, the mechanical properties of the scaffold are also critical and the modulus and elastic strength the scaffold should be similar to that of the natural bone. However, for tissue engineering applications only the mechanical properties of the final tissue engineered construct are critical (36).

TYPES OF BIOCERAMIC SCAFFOLD

Many types of porous bioceramics have been developed and are reviewed in Ref. (37). The simplest way to generate porous scaffolds from ceramics such as HA or TCP is to sinter particles. Particles are usually mixed with a wetting solution, such as poly(vinyl alcohol), and compacted by cold isostatic pressing to form a "green" body, which is sintered (heated to $\sim$1200 °C) to improve mechanical properties. Porosity can be increased by adding fillers such as sucrose to the powder and the wetting solution, which burnout on sintering. Komlev et al. (38) produced porous HA scaffolds with interconnected interparticle pore diameters of $\sim$100 μm, and a tensile strength of $\sim$0.9 MPa by sintering HA spheres 500 μm in diameter.

Other techniques include adding a combustible organic material to a ceramic powder burned away during sintering leaving closed pores; freeze drying where ice crystals are formed in ceramic slurries and then sublimation of the ice leaves pores; polymer foam replication where the ceramic slurry is poured into a polymer foam, which is then burnt out on sintering leaving a pore network. Most of these techniques produced porous ceramics that were not suitable for tissue engineering applications. Typical problems were either that the pore diameters were too low, the pores were closed, the pore distributions were very heterogeneous or mechanical strengths were very low.

Recently, rapid prototyping has been adapted for producing scaffolds with controlled and homogeneous interconnected porosity (39). Rapid prototyping is a generic term for a processing technique that produces materials in a shape determined by CAD (computer aided design) software on a computer. Such materials are usually built up layer-by-layer using a liquid phase or slurry of the material that cures or sets on contact with a substrate. Specific techniques include stereolithography, selective laser sintering, fused deposition modeling and ink-jet printing. It is a challenge to apply these techniques to direct processing of bioactive ceramic scaffolds.

Perhaps the most successful technique for synthesis of porous HA that could be produced in any size of shape, with interconnected macropore diameters in excess of 100 μm is the gel-casting process.

GEL-CASTING OF HA

In the gel-casting of HA, aqueous suspensions of HA particles, dispersing agents, and organic monomers (6 wt% acrylate/diene) are foamed. The organic monomers must be water soluble and retain a high reactivity. Foaming is the incorporation of air into a ceramic to produce a porous material. Once the slurry has foamed, *in situ* polymerization of the monomers is initiated and cross-linking occurs, forming a 3D polymeric network (gel), which produces strong green bodies. Foaming is achieved by vigorous agitation at 900 rpm with the addition of a surfactant (Tergitol TMN10; polyethylene glycol trimethylnonyl ether) under a nitrogen atmosphere (40). Surfactants are macromolecules composed of two parts, one hydrophobic and one hydrophilic. Owing to this configuration, surfactants tend to adsorb onto gas–liquid interfaces with the hydrophobic part being expelled from the solvent and a hydrophilic part remaining in contact with the liquid. This behavior lowers the surface tension of the gas–liquid interfaces, making the foam films thermodynamically stable, which would otherwise collapse in the absence of surfactant (41). Once stable bubble formation is achieved, the polymerization process is initiated using ammonium persulphate and a catalyst (TEMED, *N*,*N*,*N*′,*N*′-tetramethylethylenediamine) and the viscous foam is cast into moulds immediately prior to gelation. The surfactant stabilises the air bubbles until gelation provides permanent stability (40).

The porous green bodies are then sintered to provide mechanical strength and to burnout the organic solvents. Foam volume (and hence porosity) can be controlled by the surfactant concentration in the slurry. The materials produced exhibited interconnected pores of maximum diameter of 100–200 μm, which is ideal for tissue engineering applications.

The gel-cast HA scaffolds satisfy many of the criteria of the ideal scaffold, however, the criteria of controlled resorbability and genetic stimulation are not fulfilled. A bioactive glass scaffold would fulfil these criteria and also be able to bond to soft tissue. However, producing a 3D macroporous scaffold from a glass is difficult.

POROUS MELT-DERIVED BIOACTIVE GLASSES

Theoretically, the gel-casting process could be applied to melt-derived bioactive glass powders. However, such glasses undergo surface reactions on contact with solutions to produce an HCA surface layer and it is desirable to control the reaction before a scaffold is ready for clinical use.

Livingston et al. (42) produced a simple sintered scaffold by mixing 45S5 melt-derived bioactive glass (Bioglass) powders, with a particle size range of 38–75 μm, with 20.2 wt% camphor ($C_{10}H_{16}O$) particles, with particle size range of 210–350 μm. The mixture was dry pressed at 350 MPa and heat treated at 640 °C for 30 min. The camphor decomposed to leave porous Bioglass blocks. Macropores were in the region of 200–300 μm in diameter, however, the total porosity was just 21% as there were large distances between pores.

Yuan et al. (43) produced similar scaffolds by foaming Bioglass 45S5 powder with a dilute H_2O_2 solution and sintered at 1000 °C for 2 h to produce a porous glass–ceramic. The pores were irregular in shape and relatively few in number, implying that interconnectivity was poor, but pore diameters were in the range 100–600 μm. The pores appeared to be more like orientated channels running through the glass, rather than an interconnected network. The samples were implanted into the muscle of dogs and were found for the first time to be osteoinductive. Bone was formed directly on the solid surface and on the surface of crystal layers that formed in the inner pores. Osteogenic cells were observed to aggregate near the material surface and secrete bone matrix, which then calcified to form bone. However, although the implants had a porosity of ~30% only 3% bone was formed. It seems that creating interconnected pore networks in bioactive glasses by sintering is not practical at the present time, although sol–gel derived bioactive glasses may do so.

SOL–GEL DERIVED BIOACTIVE GLASS FOAMS: AN IDEAL SCAFFOLD?

The foaming process has also been applied to sol–gel derived bioactive glasses (44). The resulting scaffolds exhibit the majority of the criteria for an ideal scaffold.

Figure 3 shows an scanning electron microscopy (SEM) micrograph of a typical foam of the 70S30C composition (70 mol% SiO_2, 30 mol% CaO). The scaffolds have a hierarchical pore structure similar to that of trabecular bone, with interconnected macropores with diameters in excess of 100 μm and a textural porosity with diameters in range 10–20 nm (mesopores), which are inherent to the sol–gel process. The scaffolds have the potential to guide tissue growth, with the interconnected macropores providing channels for cell migration, tissue ingrowth, nutrient delivery, and eventually vascularisation (blood vessel ingrowth throughout the regenerated tissue). The mesoporous texture enhances the resorbability and biaoctivity of the scaffolds and provides nucleation points for the HCA layer and sites for cell attachment for anchorage dependant cells such as osteoblasts. The bioactive glass composition contributes high bioactivity, controlled resorbability, and the potential for the ionic dissolution products (Si and Ca) to stimulate the genes in bone cells to enhance bone regeneration.

Figure 4 shows a flow chart of the sol–gel foaming process. Sol–gel precursors [e.g., tetraethoxyl orthosilicate (TEOS, $Si(OC_2H_5)_4$] are mixed in deionized water in the presence of an acidic hydrolysis catalyst. Simultaneous hydrolysis and polycondensation reactions occur beginning with the formation of a silica network. Viscosity of the sol increases as the condensation reaction continues and the network grows. Other alkoxides–salts can be added to introduce network modifiers (e.g., CaO species). On completion of hydrolysis, the sol is foamed by vigorous agitation with the addition of a surfactant. A gelling agent [hydrofluoric acid (HF), a catalyst for polycondensation] is added to induce a rapid increase in viscosity and reduce the gelling time.

The surfactant stabilized the bubbles that were formed by air entrapment during the early stages of foaming by lowering the surface tension of the solution. As viscosity rapidly increased and the gelling point was approached, the solution was cast into airtight moulds. The gelling point is the point at which the meniscus of the foamed sol does not move, even if the mold is tilted. Casting must

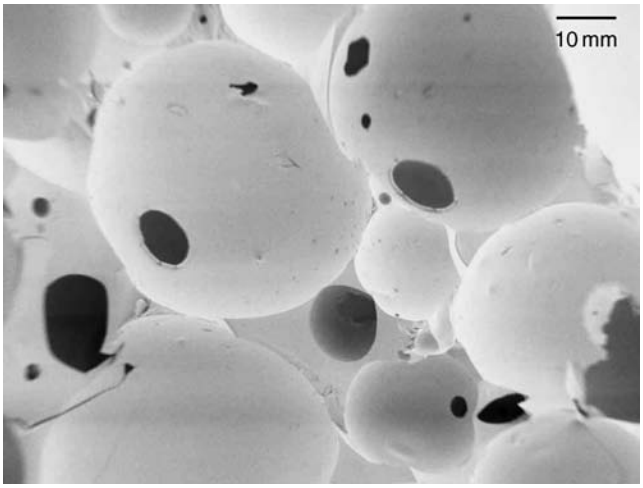


Figure 3. An SEM micrograph of a sol–gel derived bioactive glass foam scaffold.

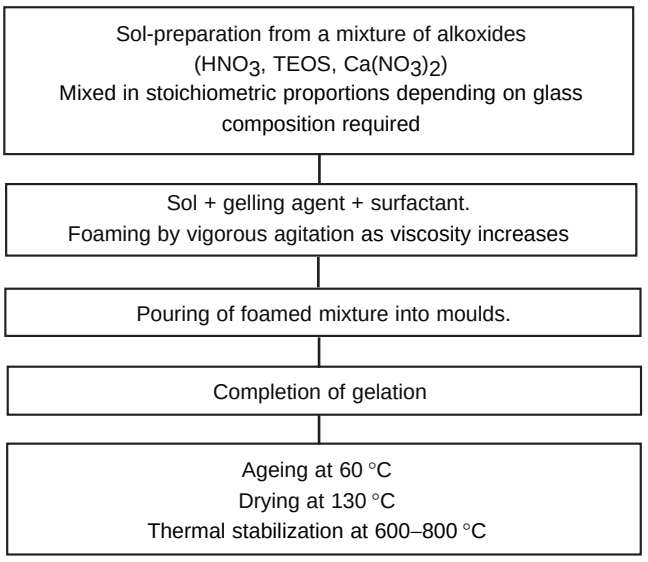


Figure 4. Flow chart of the sol–gel foaming process.

take place immediately prior to the gelation point. The gelation process provided permanent stabilization for the bubbles.

A foam scaffold is produced that sits in the liquid mixture of water and alcohol (pore liquor) produced as a byproduct of the polycondensation reaction. The foams are then subjected to the thermal treatments of ageing, drying, and stabilization. Ageing is done at 60 °C, leaving the foam immersed in its pore liquor. Ageing allows further cross-linking of the silica network and a thickening of the pore walls. Pore coarsening also occurs when larger pores grow at the expense of smaller ones.

Drying involves the evaporation of the pore liquor, which is critical and must be carried out under very carefully controlled conditions to prevent cracking under capillary pressure. Silica-based glasses that only contain the textural mesopores cannot be produced as monoliths with diameters in excess of 10 mm due to the high capillary stresses during drying. The formation of interconnected pore channels with large diameters allows efficient evaporation of the pore liquor; therefore very large crack-free scaffolds (in excess of 100 mm diameter) can be made. Thermal stabilization is carried out (again under carefully controlled heating regimes) at a minimum of 600 °C to ensure removal of silanol and nitrate groups from the glass.

The variables in each stage of the foaming process affect the final structure and properties of the foams (45,46). The percentage and pore volume of the textural mesopores can be controlled by the glass composition and the alkoxide: water ratio in initial sol preparation. Therefore the resorbability and bioactivity of the scaffolds can be easily controlled. The macropore diameters are little affected until the sintering temperature increases >800 °C. However, the glass composition, the foaming temperature, the surfactant concentration and type, the gelling agent concentration heavily affect the macropore diameters, and interconnectivity, which are vital for tissue engineering applications.

Three compositions have been successfully foamed; the tertiary 58S (60 mol% SiO_2, 36 mol% CaO, 4 mol% P_2O_5), the binary 70S30C (70 mol% SiO_2, 30 mol% CaO) composition, and 100S silica. The binary composition 70S30C (70 mol% SiO_2, 30 mol% CaO) has been found to be the most suitable to the foaming process, producing crack-free foams scaffolds with porosities in the range 60–95% (depending on the other variables in the process). Macropores were homogeneously distributed with diameters up of up to 600 μm and modal interconnected pore diameters of up to 150 μm.

Due to the nature of the sol–gel process the scaffolds can be produced in many shapes, which are determined simply by the shape of the casting mould. The scaffolds can be produced from various compositions of gel-derived glasses. All foam compositions can be easily cut to a required shape. Figure 5 shows foams produced in various shapes.

The only criterion not addressed is the matching of mechanical properties of the scaffolds to bone for *in situ* bone regeneration applications. The compressive strength of the foams (~2.5 MPa for 70S30C foams sintered at 800 °C) is less than that of trabecular bone (~10 MPa). However, the mechanical properties of these foams should be sufficient for tissue engineering applications, where bone would be grown on a scaffold in the laboratory before implantation. Work on improving the mechanical properties is ongoing.

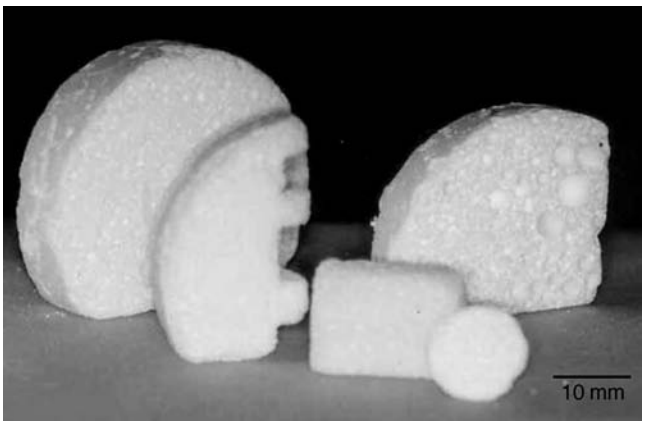


Figure 5. Sol–gel derived bioactive glass scaffolds. (Courtesy of Dr. P. Sepulveda.)

BIOLOGICAL RESPONSES TO SOL–GEL DERIVED BIOACTIVE GLASSES

The biological response to bioactive gel–glasses made from the $CaO–P_2O_5–SiO_2$ system provides evidence that bone regeneration is feasible. An important factor for future research is that the structure and chemistry of bioactive gel–glasses can be tailored at a molecular level by varying the composition (such as SiO_2 content) or the thermal or environmental processing history. The compositional range for Class A bioactive behavior is considerably extended for the bioactive gel–glasses over Class A bioactive glasses and glass–ceramics made by standard high temperature melting or hot pressing. Thus, gel–glasses offer several new degrees of freedom over the influence of cellular differentiation and tissue proliferation. This enhanced biomolecular control will be vital in developing the matrices and scaffolds for engineering of tissues and for the *in vivo* regenerative allograft stimulation of tissue repair.

Evidence of the regenerative capacity of bioactive gel–glasses powders is based on a comparison of the rates of proliferation of trabecular bone in a rabbit femoral defect (47). Melt-derived Class A 45S5 bioactive glass particles exhibit substantially greater rates of trabecular bone growth and a greater final quantity of bone than Class B synthetic HA ceramic or bioactive glass–ceramic particles. The restored trabecular bone has a morphological structure equivalent to the normal host bone after 6 weeks; however, the regenerated bone still contains some of the larger (>90 μm) bioactive glass particles. Wheeler et al. (48) showed that the use of bioactive gel–glass particles in the same animal model produces an even faster rate of trabecular bone regeneration with no residual gel–glass particles of either the 58S or 77S composition. The gel–glass particles resorb more rapidly during proliferation of

trabecular bone. Thus, the criteria of a regenerative allograft cited above appear to have been met. Recent results of *in vivo* subperiosteum implantation of 58S foams on the calvaria of New Zealand rabbits (49) showed that bone regeneration occurred more rapidly for 58S foams compared to 58S powder and that the regeneration was in line with that produced by compacted melt-derived Bioglass powders that are available commercially as Perioglas and Novabone.

SOFT TISSUE ENGINEERING

The interactions between cells and surfaces play a major biological role in cellular behavior. Cellular interactions with artificial surfaces are mediated through adsorbed proteins. A common strategy in tissue engineering is to modify the biomaterial surface selectively to interact with a cell through biomolecular recognition events. Adsorbed bioactive peptides can allow cell attachment on biomaterials, and allow 3D structures modified with these peptides to preferentially induce tissue formation consistent with the cell-type seeded, either on or within the device (50). The surface of the gel-derived foams has been modified with organic groups and proteins to create scaffolds that have potential for lung tissue engineering (50,51). If cells can recognize the proteins adsorbed on the surface of a biomaterial they can attach to it and start to differentiate, inducing tissue regeneration. However, if cells do not recognize the proteins, an immunogenic response may result, initiating a chronic inflammation that can lead to failure of the device. Besides promoting cell-surface recognition, bioactive peptides can be used to control or promote many aspects of cell physiology, such as adhesion, spreading, activation, migration, proliferation, and differentiation.

Three-dimensional scaffolds have been produced that allow the incorporation and release of biologically active proteins to stimulate cell function. Laminin was adsorbed on the textured surfaces of binary 70S30C (70 mol% SiO_2–30 mol% CaO) and ternary 58S (60 mol% SiO_2–36 mol% CaO–4 mol% P_2O_5) sol–gel derived bioactive foams. The covalent bonds between the binding sites of the protein and the ligands on the scaffolds surface do not denaturate the protein. *In vitro* studies show that the foams modified with chemical groups and coated with laminin maintained bioactivity, as demonstrated by the formation of the (HCA) layer formed on the surface of the foams on exposure to simulated body fluid (SBF). Sustained and controlled release from the scaffolds over a 30-day period was achieved. The laminin release from the bioactive foams followed the dissolution rate of the material network. These findings suggest that bioactive foams have the potential to act as scaffolds for soft tissue engineering with a controlled release of proteins that can induce tissue formation or regeneration.

The way that proteins or other bioactive peptides interact with surfaces can alter their biological functionality. In order to achieve full functionality, peptides have to adsorb specifically. They also must maintain conformation in order to remain functional biologically. Chemical groups, such as amine and mercaptan groups, are known to control the ability of surfaces to interact with proteins (51). In addition, these chemical groups can allow protein–surface interactions to occur such that the active domains of the protein can be oriented outward, where they can be maximally effective in triggering biospecific processes. Cell cultures of mouse lung epithelial cells (MLE-12) on modified 58S foam scaffolds showed that cells attached and proliferated best on 58S foam modified with amine groups (using amine propyltriethoxysilane, APTS) and coated with laminin (52).

SUMMARY

During the last century, a revolution in orthopedics occurred that has led to a remarkably improved quality of life for millions of aged patients. Specially developed bioceramics were a critical component of this revolution. However, survival of prostheses appears to be limited to ~20 years. We conclude that a shift in emphasis from replacement of tissues to regeneration of tissues should be the challenge for orthopedic materials in the new millennium. The emphasis should be on the use of materials to activate the body's own repair mechanisms, that is, regenerative allografts. This concept will combine the understanding of tissue growth at a molecular biological level with the molecular design of a new generation of bioactive scaffolds that stimulate genes to activate the proliferation and differentiation of osteoprogenitor cells and enhance rapid formation of extracellular matrix and growth of new bone *in situ*. The economic and personal benefits of *in situ* regenerative repair of the skeleton on younger patients will be profound.

BIBLIOGRAPHY

Cited References

1. Jones JR, Hench LL. Biomedical materials for the new millennium: A perspective on the future. J Mat Sci T 2001;17: 891–900.
2. Ratner BD, Hoffman AS, Schoen FJ, Lemmons JE. Biomaterials Science: An Introduction to Materials in Medicine. London: Academic Press; 1996.
3. Berry DJ, Harmsen WD, Cabanela ME, Morrey MF. Twenty-five-year survivorship of two thousand consecutive primary Charnley total hip replacements. J Bone Jt Surg 2002;84A(2): 171–177.
4. Hench LL, Polak JM. Third generation biomedical materials. Science 2002;295(5557):1014–1018.
5. Hench LL, Wilson J. An Introduction to Bioceramics. Singapore: World Scientific; 1993.
6. Hench LL. Biomaterials: A forecast for the future. Biomaterials 1998;19:1419–1423.
7. Black J, Hastings G. Handbook of Biomaterial Properties. London: Chapman and Hall; 1998.
8. Hench LL. Bioceramics. J Am Ceram 1998;81(7):1705–1728.
9. Hulbert SF. The use of alumina and zirconia in surgical implants. In: Hench LL, Wilson J, editors. An Introduction to Bioceramics. Singapore: World Scientific; 1993.
10. Hulbert SF, Bokros JC, Hench LL, Heimke G. Ceramics in Clinical Applications: Past, Present, and Future. In: Vincenzini P, editor. High tech Ceramics. Amsterdam: Elsevier; 1987.
11. Bilezikian JP, Raisz LG, Rodan GA. Principles of bone biology. London: Academic Press; 1996.

12. Sumner DR, Galante JO. Determinants of stress shielding—design versus materials versus interface. Clin Orthop Relat R 1992;274:202–212.
13. Marcus R, Feldman D, Kelsey JL. Osteroporosis. London: Academic Press; 1996.
14. de Groot K. Bioceramics of Calcium Phosphate. Boca Raton, FL: CRC Press; 1983.
15. Hench LL, West JK. Biological applications of bioactive glasses. Life Chem Rep 199;13:187–241.
16. LeGeros RZ, LeGeros JP. Dense Hydroxyapatite. In: Hench LL, Wilson J, editors. An Introduction to Bioceramics. Singapore: World Scientific; 1993.
17. Ducheyne P, Hench LL, Kagan A, Martens M, Burssens A, Mulier JC. The effect of hydroxyapatite impregnation of skeletal bonding of porous coated implants. J Biomed Mater Res 1980;14:225–237.
18. Hench LL, Splinter RJ, Allen WC, Greenlee TK. Bonding mechanism at the interface of ceramic prosthetic implants. J Biomed Mater Res 1971;74:1478–1570.
19. Sepulveda P, Jones JR, Hench LL. *In vitro* dissolution of melt-derived 45S5 and sol–gel derived 58S bioactive glasses. J Biomed Mater Res 2002;61(2):301–311.
20. Jones JR, Sepulveda P, Hench LL. Dose-dependent behaviour of bioactive glass dissolution. J Biomed Mater Res 2001;58:720–726.
21. Wallace KE, Hill RG, Pembroke JT, Brown CJ, Hatton PV. Influence of sodium oxide content on bioactive glass properties. J Mater Sci Mater Med 1999;10(12):697–701.
22. Wilson J, Douek E, Rust K. Bioglass® Middle Ear Devices: 10 Year Clinical Results. In: Hench LL, Wilson J, Greenspan DC, editors. Bioceramics 8. Oxford: Pergamon; 1995. p 239–245.
23. Fetner AE, Hartigan MS, Low SB. Periodontal repair using Perioglas® in non-human primates: Clinical and histologic observations. Comp Cont E Dent 1994;15(7):932–939.
24. Kokubo T, Ito S, Shigematsu M, Sakka S, Yamamuro T, Higashi S. Mechanical properties of a new type of apatite containing glass-ceramic for prosthetic application. J Mater Sci 1985;20:2001–2004.
25. Li R, Clark AE, Hench LL. Effect of structure and surface area on bioactive powders made by sol–gel process. In: Hench LL, West JK, editors. Chemical Processing of Advanced Materials. Vol. 56, New York: John Wiley & Sons; 1992. 627–633.
26. Hench LL, West JK. The Sol–Gel Process. Chem Rev 1990;90: 33–72.
27. Ishizaki K, Komarneni S, Nanko M. Sol–Gel Processing: Designing Porosity, Pore Size and Polarity and Shaping Processes. In: Ishizaki K, Komarneni S, Nanko M, editors. Porous Materials: Process Technology and Applications. London: Kluwer Academic Publishers; 1998. p 67–180.
28. Huang J, DiSilvo L, Wang M, Tanner KE, Bonfield W. *In vitro* mechanical and biological assessment of hydroxyapatite-reinforced polyethylene composite. J Mat S-M M 1997;8: 775–779.
29. Day R, Boccaccini AR, Roether JA, Surey S, Forbes A, Hench LL, Gabe S. The effect of Bioglass® on epithelial cell and fibroblast proliferation and incorporation into a PGA matrix. Gastroenterology 2002;122(4) T875 Suppl 1.
30. Thompson ID, Hench LL. Medical Applications of Composites. Comprehensive Composite Mater 2000; (6.39) 727–753.
31. Ohgushi H, Caplan AI. Stem Cell Technology and Bioceramics: From cell to Gene Engineering. J Biomed Mater Res B 1999;48:913–927.
32. Hench LL, Polak JM, Xynos ID, Buttery LDK. Bioactive Materials to Control Cell Cycle. Mat Res Innovat 2000;3: 313–323.
33. Xynos ID, Hukkanen MVJ, Batten JJ, Buttery LD, Hench LL, Polak JM. Bioglass® 45S5 Stimulates Osteoblast Turnover and Enhances Bone Formation In Vitro: Implications and Applications for Bone Tissue Engineering. Calciff Tiss 2000;67:321–329.
34. Xynos ID, Edgar AJ, Buttery LD, Hench LL, Polak JM. Ionic Dissolution Products of Bioactive Glass Increase Proliferation of Human Osteoblasts and Induce Insulin-like Growth Factor II mRNA Expression and Protein Synthesis, Biochem Biophys Res 2000;276:461–465.
35. Freyman TM, Yannas IV, Gibson LJ. Cellular materials as porous scaffolds for tissue engineering. Prog Mat Sci 2001;46: 273–282.
36. Holy CE, Fialkov JA, Davies JE, Shoichet MS. Use of a biomimetic strategy to engineer bone. J Biomed Mater Res 2003;65A:447–553.
37. Jones JR. Bioactive Glass 3D Scaffolds for Tissue Engineering, [dissertation]. London (UK): Imperial College London; 2002.
38. Komlev VS, Barimov SM. Porous hydroxyapaptite ceramics of bi-modal pore size distribution. J Mater Sci Mater Med 2002;13:295–299.
39. Chu GTM, Orton DG, Hollister SJ, Feinberg SE, Halloran JW. Mechanical and *in vivo* performance of hydroxyapatite implants with controlled architectures. Biomaterials 2002; 23:1283–1293.
40. Sepulveda P, Binner JGP, Rogero SO, Higa OZ, Bressiani JC. Production of porous hydroxyapatite by the gel-casting of foams and cytotoxic evaluation. J Biomed Mater Res 2000;50: 27–34.
41. Rosen MJ. Surfactants and Interfacial Phenomena. 2nd ed, New York: Wiley; 1989. p 277–303.
42. Livingston T, Ducheyne P, Garino J. *In vivo* evaluation of a bioactive scaffold for bone tissue engineering. J Biomed Mater Res 2002;62:1–13.
43. Yuan H, de Bruijn JD, Zhang X, van Blitterswijk CA, de Groot K. Bone Induction by porous glass ceramic made from Bioglass® (45S5). J Biomed Mater Res 2001;58(3): 270–276.
44. Sepulveda P, Jones JR, Hench LL, Bioactive sol–gel foams for tissue repair. J Biomed Mater Res 2002;59(2):340–348.
45. Jones JR, Hench LL. The effect of processing variables on the properties of bioactive glass foams. J Biomed Mater Res In press.
46. Jones JR, Hench LL. The effect of surfactant concentration and glass composition on the structure and properties of bioactive foam scaffolds. J Mat Sci In press.
47. Oonishi H, Hench LL, Wilson J, Sugihara F, Tsuji E, Kushitani S, Iwaki H. Comparative bone growth behaviour in granules of bioceramic materials of various sizes. J Biomed Mater Res 1999;44(1):31–43.
48. Wheeler DL, Hoellrich RG, McLoughlin SW, Chamerland DL, Stokes KE. In Vivo Evaluation of Sol–Gel Bioglass®–Biomechanical Findings. In: Sedel L, Rey C, editors. Bioceramics. Volume 10, 1997. p 349–350.
49. Cook R 58S sol–gel Bioglass: a study of osteoproductive, interfacial and handling properties using new microscopic techniques. [dissertation] London (UK). University of London; 2003.
50. Lenza RFS, Jones JR, Vasconcelos WL, Hench LL. *In vitro* release kinetics of proteins from bioactive foams. J Biomed Mater Res In press.
51. Lenza RFS, Jones JR, Vasconcelos WL, Hench LL. *In vitro* release kinetics of proteins from bioactive foams. J Biomed Mater Res In press.
52. Mansur HS, Vasconcelos WL, Lenza RFS, Oréfice RL, Reis EF, Lobato ZP. Sol–gel silica based networks with controlled chemical properties. J Non-Cryst 2000;273:109–115.
53. Tan A, Romanska HM, Lenza R, Jones J, Hench LL, Polak JM, Bishop AE. The effect of 58S bioactive glass sol–gel

derived foams on the growth of murine lung epithelial cells. Key Eng Mat 2003;240–242: 719–724.

References List

Clifford A, Hill R, Rafferty A, Mooney P, Wood D, Samuneva B, Matsuya S. The influence of calcium to phosphate ratio on the nucleation and crystallization of apatite glass-ceramics. J Mater Sci Mater Med 2001;12(5): 461–469.

Healy KE. Molecular engineering of materials for bioreactivity. Curr Op Sol 1999;4: 381–387.

See also BIOMATERIALS FOR DENTISTRY; BONE AND TEETH, PROPERTIES OF; HEART VALVE PROSTHESES; HIP JOINTS, ARTIFICIAL.

BIOMATERIALS: CARBON

ROBERT B MORE
RBMore Associates,
Austin, Texas

JACK C BOKROS
Medical Carbon Research Institute
Austin, Texas

INTRODUCTION

Inorganic, elemental carbon is one of the oldest, and yet newest, biomaterials. Carbon utilization began with prehistoric human's use of charcoal and continues today with a variety of applications exploiting the physicochemical, adsorptive, structural, and biocompatible properties of different forms of carbon. To date, the most important carbon biomaterials have been the isotropic pyrolytic carbons (PyC), produced in a fluidized bed, for use as structural blood contacting components for heart valve prostheses and for small joint orthopedic prostheses. Adsorptive properties of activated carbons also find widespread use for the removal of toxins from the body either by direct ingestion, dialysis, or by plasmapherisis.

Other carbons, such as carbon fibers and glassy carbons have been proposed for use in a variety of structural implants, but because of limited strength and durability, have not been generally accepted. However, carbon fibers and glassy carbons are used as electrodes and electronic components in biomedical analytical devices. Diamond-like coatings have been proposed to provide enhanced wear resistance for large orthopedic components, but this technology is still under development. For the future, carbon holds a central focus in nanotechnology with investigations into the use of fullerenes and carbon nanotubes as means of imaging and manipulating nanoscale bioactive molecules, as selective markers, and perhaps as inhibitors to virulent organisms such as the human immunodeficiency virus (HIV).

Elemental carbon is allotropic, meaning that it can exist in two or more forms (1). There are at least two perfectly crystalline allotropic forms: graphite and diamond, and a myriad of intermediate, imperfectly crystalline, amorphous structures (2). This diversity in structure leads to considerable variability in physical and mechanical properties ranging from graphite, one of the softest materials, to diamond, the hardest material known to human. Thus, carbon rather than being a single material is actually a spectrum of materials (3). For this reason, it is necessary to qualify the use of the term *carbon* as designating a generic material with a carbon elemental composition. A specific carbon material must then be qualified with a description of it's structure.

In general, most of the pure carbons are biocompatible in that they are bioinert, do not provoke thrombosis, hemolysis, inflammatory response, nor activate the complement system (4). Furthermore pure carbons are biostable: toxic products are not generated and the materials retain their properties. However, just because a candidate material is a carbon does not mean that its particular microstructure and properties are appropriate for the desired application. For example, structural applications such as cardiovascular and orthopedic prostheses require strength, fatigue resistance, wear resistance, low friction and durability, in addition to tissue compatibility (3). Not all carbons have the appropriate properties needed for structural use.

In order to appreciate the medically important carbons, some of the various forms of elemental carbon, their synthesis, structure, and properties will be presented and briefly discussed. We will then return to the important carbon biomaterials, discuss their utilization, and conclude with speculations as future directions.

BACKGROUND

Structure of Carbons

Diversity in carbon arises from the electronic configuration: $1s^2 2s^2 2p^2; {}^3P$, which allows the formation of a number of hybridized atomic orbitals that share four valence electrons to form covalent bonds with directional properties. On the basis of bond structures that arise from the hybridized orbital bonds, carbon compounds are classed as aliphatic and as aromatic (5). Originally, aliphatic meant "fatty" and aromatic meant "fragrant", but these descriptions no longer have any real significance. Aliphatic compounds are further subdivided into alkane, alkenes, alkynes, and cyclic aliphatic. Aromatic compounds are benzenes and compounds that resemble benzene in chemical behavior. With a few exceptions, organic compounds of medical importance tend to be aromatic or benzene-like. Details of electronic structure beyond that given here may be found in standard chemistry and organic chemistry textbooks (1,5).

Naturally Occurring Carbons

Diamond. Diamond is the ultimate polycyclic aliphatic system, but is not a hydrocarbon; rather, it is one of the allotropic forms of elemental carbon. In the diamond allotropic structure, one s and three p orbitals undergo hybridization to form the sp^3 orbital that has tetrahedral symmetry. This symmetry allows covalent bonds to connect each carbon atom with four others. Bond angles are 109.5° and the carbon–carbon bond length is 0.154 nm. Each carbon is bonded to the original plus three others and this structure propagates throughout the entire crystal forming one giant isotropic molecule of covalently bonded carbons (1,2), as shown in Fig. 1. The diamond crystallographic

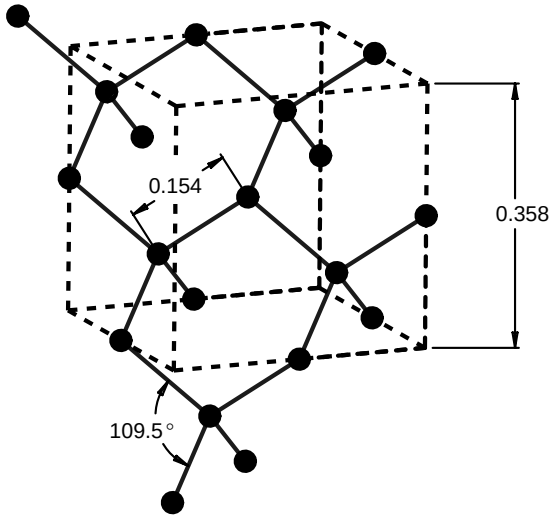


Figure 1. Crystallographic structure of diamond with tetrahedral bond angles of 109.5° and bond lengths of 0.154 nm. The unit cell with a length of 0.358 nm is shown by the dashed lines. The spheres represent the location of the atoms and not size.

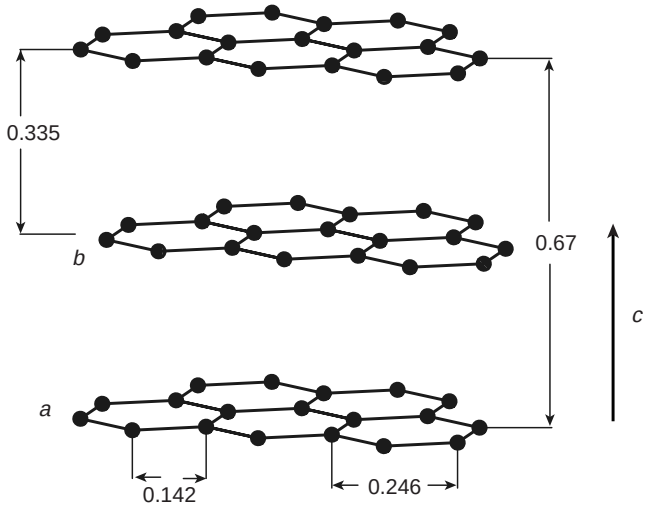


Figure 2. Crystallographic structure of graphite. Basal planes *a* and *b* contain the hexagonal covalently bonded carbons with bond angles of 120° and bond lengths of 0.142 nm. Because of sp^2 coordination, each basal plane is shifted one atomic position relative to one another. The successive basal planes are separated by 0.34 nm in the *c* direction. The distances 0.246 and 0.67 nm are the dimensions of the graphite hexagonal close-packed unit cell.

structure can be visualized as a repetition of the six-carbon cyclohexane "chair" configuration. Because of the large number covalent bonds with an interlocking isotropic orientation, the structure is very rigid. A large amount of energy is required to deform the crystal, hence diamond is the hardest material known.

Graphite. Where diamond is the ultimate polycyclic aliphatic system, graphite is the ultimate polycyclic aromatic system. Graphite has a layered structure consisting of planar arrays in which each carbon atom is bonded by two single bonds and one double bond with its three nearest neighbors. Where diamond can be visualized as a repeated cyclohexane chair, graphite is visualized as a repeated six-carbon benzene ring structure. Within a single plane, each carbon is bonded with a single atomic distance of 0.142 nm to its three nearest neighbors by sp^2 orbitals with hexagonal symmetry and 120° bond angles (1). Three of the four valence electrons are used to form these regular covalent σ (sigma) bonds, which forms the basal planes of hexagonal covalently bonded atoms as shown in Fig. 2. A single basal layer of the hexagonal carbons is known as a *graphene* structure.

The fourth π (pi) electron resonates between valence structures in overlapping *p* orbitals forming π bond donut-shaped electron clouds with one lying above and one below and perpendicular to the plane of the σ bonded carbons (2). Successive layers of the hexagonal carbons are held together at a distance of 0.34 nm by weak van der Waals forces or by interactions between the π orbitals of the adjacent layers (6,7). Thus the graphite structure is highly anisotropic, consisting of stacked parallel planes of strong covalent in-plane bonded carbons with the planes held together by much weaker van der Waals type forces. Because of weak interlayer forces, the layers are easily separated, which accounts for softness and lubricity of graphite. These weak interlayer forces also account for (a) the tendency of graphitic materials to fracture along planes, (b) the formation of interstitial compounds; and (c) the lubricating, compressive, and many other properties of graphite (2,6).

Amorphous Carbons. There are many cystallographically disordered forms of carbon with structures that are intermediate between those of graphite and diamond. The majority tends to be imperfectly layered graphene, turbostratic, and randomly oriented structures (2). X-ray diffraction patterns for amorphous carbons are broad and diffuse because of the small crystallite size, imperfections, and a turbostratic structure (2). In turbostratic structures, there is order within the graphene planes (denoted as *a* and *b*), but no order between planes (denoted as *c* direction) as shown in Fig. 3. Crystallographic defects such as lattice vacancies, kinked or warped layer planes, and possible aliphatic bonds tend to increase the turbostratic layer spacing relative to graphite and inhibit the ability of the layer planes to slip easily as occurs in graphite (2). Like graphite, there is strong in plane covalent bonding, but, because the ability of the planes to slip past one another is inhibited, the materials are much harder and stronger than graphite. Turbostratic carbons occur in a spectrum of amorphous ranging through mixed-amorphous structures and include materials such as soot, pitch, and coals.

Fullerenes. The recently discovered fullerenes (2,8,9) can occur naturally as a constituent of soot. Fullerenes are hollow cage-like structures that can be imagined as graphene sheets that have been folded or rolled into a ball or cylindrical tube. However, the structures are actually formed by the reassociation of individual carbon atoms rather than a folding or rolling of a graphene structure. The most famous fullerene is the ~1 nm diameter C_{60}

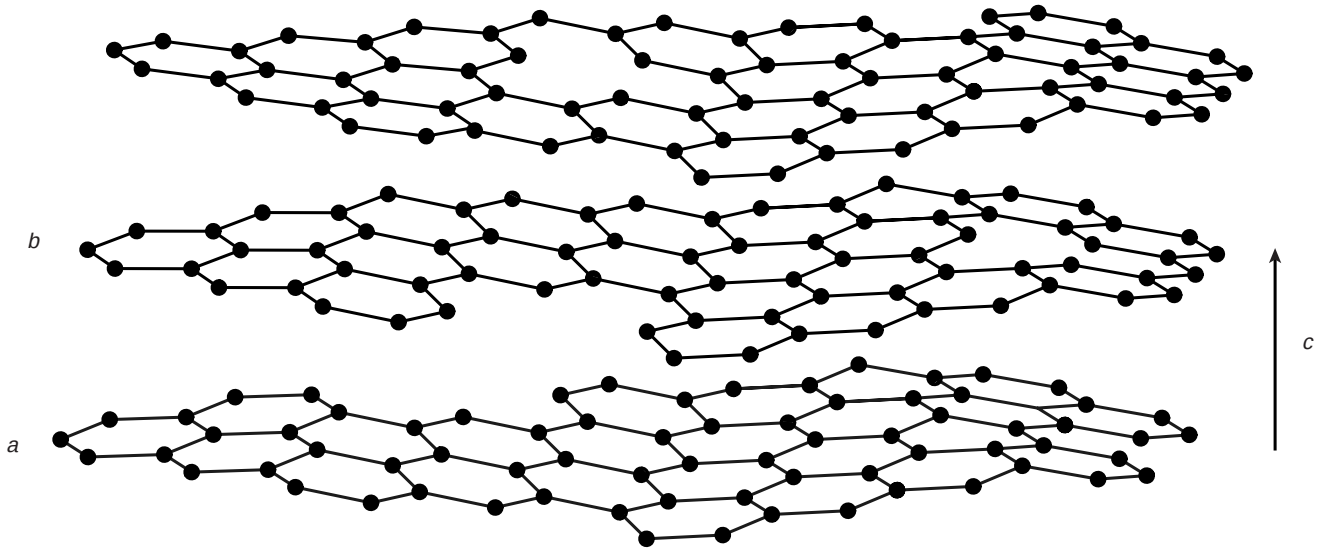


Figure 3. Turbostratic amorphous structure.

(60 carbon) buckministerfullerene (bucky ball) with a truncated icosahedron structure that resembles a European football. Because the structure is reminiscent of the geodesic dome designed by the architect Buckminister Fuller, the proposed structure was named after him (8).

Geometrically, the bucky ball has a repeating structure that consists of a pentagon surrounded by five hexagons (see Fig. 4). In order to wrap into a nonplanar ball, the graphitic p orbitals must assume an angle of 101.6° relative to the plane of the C bonds rather than 90° for graphite (9). There are a number of other possible carbon number ball structures, but the smallest sizes are thought to be limited to C_{60} and C_{70} by the molecular strain induced at the edge-sharing pentagons. Although remarkably stable, C_{60} can photodisassociate when pulsed with laser light and loose carbon C_2 pairs down to $\sim C_{32}$, where it explodes into open fragments because of strain energy (10).

Metals can also be inserted into the buckyball cage simply by conducting fullerene synthesis in the presence of metals (11). Such internally substituted endohedral fullerenes are fancifully called "dopyballs" for doped fullerenes (12) and denoted as M_aC_n, where M_a is the metal and C_n the carbon complex. "Fullerite" refers to a solid-state association of individual C_{60} molecules, named by analogy to graphite, in which the bucky balls assume a face-centered cubic (fcc) crystallographic structure with lattice constant $a = 1.417$ nm (13). Treatment of fullerite with 3 equiv of alkali metal, A_3C_{60}, makes it a super conductor at room temperature (14), whereas treatment with 6 equiv of alkali metal, A_6C_{60}, makes it an insulator.

An excellent introduction to fullerenes by Bleeke and Frey, Department of Chemistry, Washington University, St. Louis, MO, is available on the Internet at http://www.chemistry.wustl.edu/~edudev/Fullerene/fullerene.html (15).

Nanotubes. Although most likely synthetic, because of the basic fullerene structure, nanotubes will be discussed here. A nanotube consists of a single graphene sheet SWNT (single-wall nanotube) or multiple concentric graphene sheets MWNT (multiwall nanotube) rolled into a cylindrical tube (16). In MWNT, the nested concentric cylinders are separated by the ~ 0.34–0.36 nm graphite layer separation distance.

There are several different wrapping symmetries to give chiral, zigzag or arm chair nanotubes and the tubes may be end capped by a bucky ball half-sphere. Lengths range from well > 1 μm and diameters range from 1 nm for SWNT to 50 nm for MWNT. A zigzag SWNT is shown in Fig. 5. Additional information regarding nanotubes can be found at Tomanek's laboratory, at the University of Michigan. A very informative web page dedicated to nanotubes (17) is at http://www.pa.msu.edu/cmp/csc/nanotube.html.

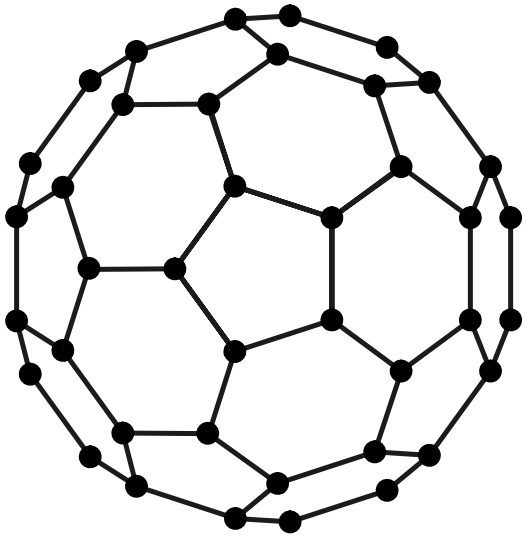


Figure 4. A surface view of a C_{60} structure, buckministerfullerene (buckyball), with an ~1 nm diameter, is shown.

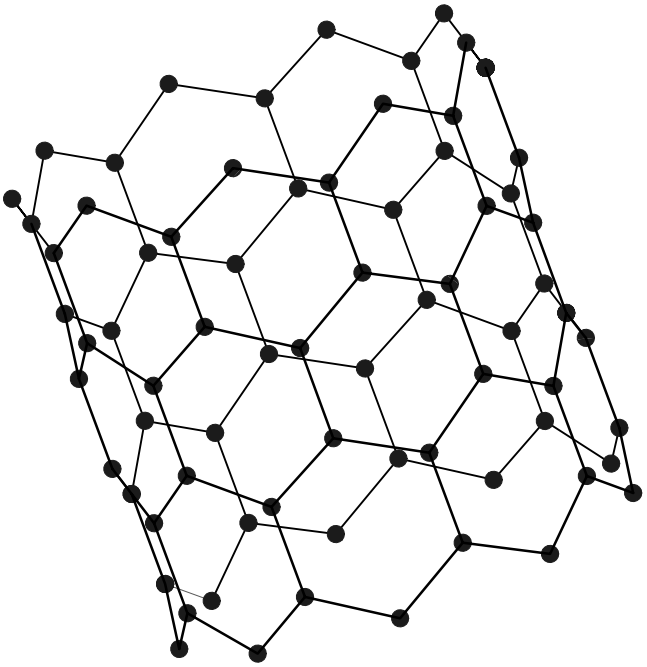


Figure 5. A short section of SWNT with a zigzag chiral symmetry is shown. The arrow indicates the long axis of the tube and bonds on the forward wall are more heavily drawn. Like the buckyball, this SWNT has a diameter of ~1 nm.

Synthetic Carbons

Carbon structures can be synthesized through a variety of processes. Because these processes define the resulting materials, both will be presented together. The most important synthesis processes include carbonization or pyrolysis and graphitization (2). Carbonization is a thermal process in which an organic precursor is converted into an all carbon residue with the diffusion of non-carbon volatile compounds to the atmosphere (2,6). The resulting all-carbon residue is known as a coke or a char.

Coke is a graphitizable carbon and chars are nongraphitizing (2). Cokes and chars are amorphous, lacking long-range crystallographic structure (turbostratic), with the degree of structure dependant on the precursor and the particular carbonization process. A coke may then be graphitized. In graphitization, residual non-carbon impurities are removed and the turbostratic structure is converted into a well-ordered graphite crystallographic structure by heating to high temperatures (6). A char, when graphitized retains its disordered turbostratic structure (2).

Synthetic Graphites. These carbons are prepared by grinding or milling a solid precursor material (coke) into fine particles, binding with a material such as coal tar pitch, and then molding into shape (2,6). The resulting material is then carbonized baked and graphitized. Typical milled grain sizes may range from 1 μm up to ~1 cm. The mixture of filler and binding may be doped or impregnated with the addition of non-carbon elements such as tungsten. The final properties of the molded graphite depend on the degree of graphitization and the grain size (6). Other important parameters are porosity, anisotropy, and density (6). One synthetic graphite used in medical devices, POCO AXF 5Q, has a grain size of 5 μm, a pore size of 0.6 μm, and a 23% volume total porosity. This particular graphite grade is often mixed with 10% by weight fine powdered tungsten before molding and baking to confer radio opacity (6).

Viterous, Glassy Carbons. Carbonization of certain polymer chars produces glassy carbons. These materials are amorphous, turbostratic, and thought to contain some sp^3 character in addition to sp^2(2). Precursors are polymers such as phenolformaldehyde, poly(furfuryl alcohol) and poly(vinyl alcohol). Shapes are attained by carbonization in molds and are limited to ~7 mm thickness because of volumetric shrinkage (~50%) and the need for gases generated during carbonization to diffuse out and not nucleate bubbles (18). The resulting material is hard, brittle, and difficult to machine.

Carbon Fibers. Thomas Edison produced the first carbon fiber in 1879 as a filament of an incandescent lamp and the first patent was issued in 1892 (2,3). Hiram Maxim received a process patent for the production of carbon fibers in 1899 (3). However, prior to the 1950s carbon fibers were of marginal strength and used primarily for their electrical properties.

Carbon fibers are highly oriented, small (with diameters on the order of 7 μm), crystalline filaments that are prepared by carbonization of polymeric filament precursors and sequential heat treatment. There are three classes of fibers based on the precursor material: PAN (polyacrylonitrile), rayon, and pitch (2). Other precursors and processes exist, but have not been as successful commercially (3). In general, fibers are classified according to structure and degree of crystallite orientation (2): high modulus (345 GPa and above), intermediate modulus (275 GPa), and low modulus (205 GPa and below). Structures are turbostratic and can contain mixed sp^2 and sp^3 bonding (2). Because of their small volume, tensile strengths can be quite high, on the order of 1000–7000 MPa.

Chemical Vapor Deposited (CVD) Carbons. Carbonization of a gaseous or liquid precursor such as gaseous hydrocarbons produces a material known as pyrolytic carbon or pyrolytic graphite (2). Thermal decomposition of hydrocarbons produces carbon free radicals in the vapor phase, which can then polymerize to form coatings on exposed surfaces. Common precursor hydrocarbons are methane, propane, and acetylene. The resulting turbostratic pyrolytic carbons can be isotropic or anisotropic depending on the pyrolysis reaction conditions (19). The coating process can be prolonged so as to produce structural components for heart valve and orthopedic prostheses with coating thickness on the order of 1 mm.

The pyrolytic carbons for medical applications are formed by CVD processes in fluidized-bed reactors (20). Propane is the precursor gas and an inert gas such as nitrogen or helium provides the levitation needed to fluidize a bed of small refractory particles. Graphite preformed substrates (e.g., POCO AXF 5Q) suspended in the fluidized

bed are coated with the pyrolytic carbon (20). The resulting coating structures are turbostratic and isotropic with very small randomly oriented crystallites: These crystallites will henceforth be designated as *isotropic fluidized-bed pyrolytic carbons*. Nonfluidized-bed CVD reactors tend to produce a highly anisotropic coating with column-like, (columnar) crystallites or laminar structures with the basal planes oriented parallel to the deposition surface (2,19).

Highly Oriented Pyrolytic Graphite (HOPG). Columnar and laminar pyrolytic carbons when annealed $>2700\,^{\circ}C$ are reordered, the turbostratic imperfections disappear and the resulting structure closely approaches the ideal graphite structure with an angular spread of the crystallite *c* axes of $<1^{\circ}$ (2).

Vapor-Phase Carbons. Carbon CVD coatings formed from solid precursors carbonized by vaporization are considered vapor-deposited coatings (VPC). Precursors can be graphite or vitreous carbon vaporized by heating to high temperature at low pressure to generate the carbon free radicals. This technique produces line-of-sight coatings of nanometer and micrometer level thickness. The VPC coatings tend to be turbostratic and amorphous (3).

Diamond-Like Carbon. Diamond-like carbon coatings containing mixed sp^3 and sp^2 bonds can be prepared by physical vapor deposition (PVD). These PVD methods produce carbon free radicals by ion beam sputtering, or laser or glow discharge of solid carbon targets. There are also mixed PVD/CVD methods such as plasma or ion beam deposition from hydrocarbon gas precursors (2).

Activation. Activated carbons have large surface areas and large pore volumes that lend to a unique adsorption capability (21). Activation is a thermal or chemical treatment that increases adsorption capability. The mechanisms for adsorption are complex and include physical and chemical interactions between the carbon surface and the sorbed substances. Activity includes (a) adsorption, (b) mechanical filtration, (c) ion exchange, and (d) surface oxidation (22). Of these, adsorption and surface oxidation are the most important for medical applications. Incompletely bonded basal plane carbons as occur at crystal edges exposed at the surface, as well as defects, are chemically active and can chemisorb substances, particularly oxidizing gases such as carbon monoxide and carbon dioxide (23). Surface oxidation involves the chemisorbance of atmospheric oxygen and further reaction of the sorbed oxygen with other substances (24). Physical adsorbance occurs because of charge interactions, and chemical adsorbance occurs because of reactions between the adsorbant and adsorbate (24).

Any high carbon material can be "activated" by various oxidizing thermal and chemical processes that increase porosity and active surface area, which increases the ability for chemisorption (25). A char is formed and then treated chemically or physically to generate pores and the surface oxidized (21). Surface oxide complexes such as phenols, lactones, carbonyls, carboxylic acids, and quinones form that have a strong affinity for adsorbing many substances such as toxins or impurities (26). Carbon fibers may be activated in order to enhance the ability to bind with a matrix material when used as a filler.

PROPERTIES

Representative physical and mechanical properties of the carbon allotropes are summarized in Table 1 (2,3,27). Materials included span the spectrum from natural diamond to natural graphite. There is considerable variability

Table 1. Representative Mechanical and Physical Properties for Carbon Allotrophs

Property	Natural Diamond	Amorphous Carbons	HOPG	Natural Graphite
Density, $g \cdot cm^{-3}$	3.5–3.53	1.45–2.1	2.25–2.65	2.25
Young's modulus, GPa	700–1200	17–31	20	4.8
Hardness, mohs	10	2–5		1–2
Hardness, DPH 500 g		150–(>230)		
Flexural strength, MPa		175–520	80 (*c*) 120 (*ab*)	
Compressive yield strength, MPa	8680–16530	700–900	100	
Fracture toughness, $MPa \cdot m^{1/2}$	3.4	0.5–1.67		
Poisson's ratio	0.1–0.29	0.2–0.28		
Wear resistance	Excellent	Poor to excellent	Poor	Poor
Electrical resistivity, Ω·cm	2700		0.15–0.25 (*c*) 3.5×10^{-5}–4.5×10^{-5}(*ab*)	0.006
Magnetic susceptibility, $\times 10^6$ emu/mol	−5.9			−6
Melting point, °C	3550		3650	3652–3697 (sublimes)
Boiling point, °C	4827			4220
CTE linear, $(20^{\circ}C)\mu m \cdot (m \cdot {}^{\circ}C)^{-1}$	1.18	2.6–6.5	−0.1 (*ab*) 20 (*c*)	0.6 (*ab*) 4.3 (*c*)
Heat capacity, J/g·°C	0.4715			0.69
Thermal conductivity, $W \cdot (m \cdot K)^{-1}$	2000	4.6–6.3	16–20 (*ab*) 0.8 (*c*)	19.6 (*ab*) 0.0573 (*c*)

[a]Values from Matweb.com and from Refs. (2,3).

in properties depending on the structure, anisotropy, and crystallinity, particularly in the amorphous carbons. Physical properties such as resistivity, coefficient of thermal expansion, thermal conductivity, and tensile strength (28) show profound sensitivity to direction in the graphitic materials. This anisotropy is most easily seen in HOPG by comparing the *ab* direction, parallel to the σ-bonded basal plane, to the perpendicular *c* direction. For example, the resistivity drops for HOPG because of the mobility of the π-electron clouds in the *ab* direction relative to the *c* direction (2). Diamond, with full covalent bonding, is an insulator.

Thermal conductivity, which occurs by lattice vibration, is related to a mean-free-path length for wave scattering. Little scattering occurs in the near-perfect graphite crystal basal plane, so the scattering path length and thermal conductivity are high in the *ab* direction. In the *c* direction, thermal conductivity is much lower because the amplitude of lattice vibration is considerably lower than for the *ab* direction (2). Thermal expansion is related to the interatomic spacing of the carbon atoms, bond strength, and vibration. As temperature increases, the atoms vibrate and the mean interatomic spacing increases. For weak bonding in the *c* direction, the interatomic vibrational amplitude and dimensional changes are larger than for the strongly bonded *ab* direction (2). The CTE values are stated for room temperature to $\sim 200\,^{\circ}C$; the negative values shown in Table 1 are possibly due to internal stresses and become positive at higher temperatures. Large anisotropic differences in CTE can result in large internal stresses and possible structural problems with heating and cooling over large temperature differences.

BIOCOMPATIBILITY

Pyrolytic carbons used in heart valve and orthopedic prostheses have a successful clinical experience as long-term implant materials for blood and skeletal tissue contact (3,29–31). These isotropic, fluidized-bed, pyrolytic carbons that were originated at General Atomics in the 1960s demonstrate negligible reactions in the standard Tripartite and ISO 10993-1 type biocompatibility tests. Results from such tests are given below in Table 2 (20). This material is so nonreactive that it has been proposed as a negative control for these tests. However, the surface is not totally inert and is capable of adsorption and desorption of a variety of substances including protein (32–39). The mechanism for biocompatibility is yet poorly understood, but probably consists of a complex, interdependent, and time-dependent series of interactions between the proteins and the carbon surface (32).

Because of the similarity in surface sp^2 and sp^3 character among the various pure carbons, most can be expected to have the tissue compatibility and biostability to perform well in these biocompatibility tests also. Vitreous carbons (40), activated carbons, and diamond-like coatings (41) are known to exhibit tissue compatibility, likewise the fullerenes will probably be found tissue compatible. As an extreme example, in testing the safety of an activated charcoal for hemoperfusion, Hill (42) introduced finely ground charcoal suspensions into the blood stream of rats in varying concentrations up to 20 mg/kg charcoal and observed no differences in survival or growth relative to controls over a 2-year observation period.

A reasonable working definition for biocompatibility has been given by Williams (43) as, "*The ability of a material to perform with an appropriate response in a specific application*". The important point here is that while many carbons provoke a minimal biological reaction, "*the specific application*" demands a complete set of mechanical and physical properties, in addition to basic cell compatibility. Because there are a number of possible microstructures, each with different properties, a given carbon will probably not have the entire set of properties needed for a specific application. Historically, the clinically successful isotropic, fluidized-bed, pyrolytic carbons required extensive development and tailoring to achieve the set of mechanical and physical properties needed for long-term cardiovascular and orthopedic applications (20,30–32).

Blood compatible glassy carbons, for example, are often proposed for use in heart valves. However, glassy carbons were evaluated in the early 1970s as a replacement for the polymer Delrin in Bjork–Shiley valve occluders and actually found to have inferior wear resistance and durability relative to the polymer (44). Thus, the fact that a material is carbon, a turbostratic carbon, or a pyrolytic carbon and is cell compatible, does not justify its use in a long-term implant devices (3,32,33). The entire range of physical and mechanical properties as dictated by the intended application are required.

MEDICAL APPLICATIONS

Activated Charcoal–Activated Carbons

Charcoal, the residue from burnt organic matter, was probably one of the first materials used for medical and

Table 2. Biological Testing of Pure PyC

Test description	Protocol	Results
Klingman maximization	ISO/CD 10993-10	Grade 1; not significant
Rabit pyrogen	ISO/ DIS 10993-11	Nonpyrogenic
Intracutaneous injection	ISO 10993-10	Negligible irritant
Systemic injection	ANSI/AAMI/ISO 10993-11	Negative—same as controls
Salmonella typhimurium Reverse mutation assay	ISO 10993-3	Nonmutagenic
Physiochemical	USP XXIII, 1995	Exceeds standards
Hemolysis–rabbit blood	ISO 10993-4/NIH 77-1294	Nonhemolytic
Elution test, L929 mammalian cell culture	ISO 10993-5, USP XXIII, 1995	Noncytotoxic

biocompatible applications. Prehistoric humans knew that pulverized charcoal could be placed under the skin indefinitely without ill effects, thus allowing decorative tattoos (45). Because granulated charcoal has an active surface area, it can adsorb toxins when ingested. Likewise, charcoal has long been used to clear water and other foods. The ancient Egyptians first recorded the medical use of charcoal $\sim$1500 BC (21). During the 1800s, the first formal scientific studies of charcoal as an antidote to treat human poisoning appeared in Europe and The United States. In some of these studies, the researchers demonstrated charcoals effectiveness by personally ingesting charcoal along with an otherwise fatal dose of strychnine or arsenic (21). Activation was discovered $\sim$1900 and activated charcoals were used as the sorbant in World War I gas masks (21).

Today's activated carbons or activated charcoals are derived from a number of precursor organic materials ranging from coal, wood, coconut shells, and bone. Chars are prepared by pyrolyzing the starting organic material using heat in the absence of oxygen. The char is then activated by treatment with chemicals or steam. Activated carbon has remarkable adsorptive properties that vary with the starting material and activation process. Common active surface areas are on the order of 1000–2000 m^2/g. Prior to the discovery of activation processes, charcoals were naturally oxidized by exposure to the atmosphere and moisture, as in charcoal, or oxidized in a more controlled activating process (46).

Orally administered activated carbon applications include use as an antidote to poisoning and to drug overdoses, where it acts at the primary site of drug adsorption in the small intestine. There are no contraindications for patients with intact GI tracts. There are numerous advantages and few disadvantages. One of the main disadvantages is that it is unpleasant for the health care professional to use because it can be messy, staining walls, floors, clothing, and so on. It may also be unpleasant to swallow because of a gritty texture (46).

There are extracorporal, parenteral, methods such as hemoperfusion, hemofilteration, and plasmapheresis where activated carbon is used to remove toxins from a patient's blood. The patient's heparinized blood is passed via an arterial outflow catheter into an extracorporal filter cartridge containing the activated carbon and then returned to the patient via a venous catheter. These techniques are effective when there is laboratory confirmation of lethal toxin concentrations in the blood and for poorly dialyzable and nondialyzable substances (47).

Pyrolytic Carbons

Isotropic, fluidized-bed PyCs, appropriate for cardiovascular applications originated at General Atomics in the late 1960s as a cooperative effort between an engineer, Jack Bokros, working with pyrolytic carbons as coatings for nuclear fuel particles and a surgeon, Vincent Gott, who was searching for thromboresistant materials for cardiovascular applications (48). Together, they tailored a specific fluidized-bed, isotropic pyrolytic carbon alloy with the biocompatibility, strength and durability needed for long-term structural applications in the cardiovascular system. The original material was a patented silicon-alloyed pyrolytic carbon given the tradename "Pyrolite" (20).

In the early 1960s, heart valve prostheses constructed from polymers and metal were prone to early failure from wear, thrombosis, and reactions with the biological environment. Prosthesis lifetimes were limited to several years because of wear in one or more of the valve components. Incorporation of PyC as a replacement for the polymeric valve components successfully eliminated wear as an early failure mechanism. Subsequently, in most valve designs, metallic materials were replaced with PyC also (20,29–33,49).

During the 1970s and 1980s Pyrolite was only available from a single source until the original patents expired. Since that time, several other sources have appeared with copies of the original silicon-alloyed General Atomics material. In the early 1990s, the Bokros group revisited the synthesis methods and found that with the then available technology for process control, that a pure carbon pyrolytic carbon could be made with better mechanical properties and potentially better biocompatibility than the original silicon-alloyed Pyrolite (20). This new pure isotropic, fluidized-bed, pyrolytic carbon material was patented and named On-X carbon. On-X carbon is currently utilized in mechanical heart valves and in small joint orthopedic applications.

These PyC materials are turbostratic in structure and isotropic with fine randomly oriented crystallite sizes on the order 2.5–4.0 nm and c layer spacing of $\sim$0.348 nm (50–52). Implants are prepared by depositing the hydrocarbon gas precursor coating in a fluidized bed on to a preformed graphite substrate to a thickness of $\sim$0.5 mm (29–32,53). The coatings then may be ground and polished if desired and subjected to a proprietary process that minimizes the degree of surface chemisorbed oxygen.

Some of the mechanical and physical properties of the pure and silicon-alloyed PyC materials appropriate for use in long-term implants are given Table 3 (3,20). A typical glassy carbon and a fine-grained synthetic graphite are also included for comparison. The PyC flexural strength, fatigue, and wear resistance provide adequate structural integrity for a variety of implant applications. The density is low enough to allow components to be actuated by flowing blood. Relative to orthopedic applications, Young's modulus is in the range reported for bone (54,55), which allows for compliance matching and minimizes stress shielding at the prosthesis bone interface. The PyC strain-to-failure is low relative to ductile metals and polymers; but it is high relative to ceramics. Because PyC is a nearly ideal linear elastic material, component design requires the consideration of brittle material design principals. Certain properties such as strength vary with the effective stressed volume, or stressed area as predicted by Weibull theory (56). Table 3 strength levels were measured for specimens tested in four-point bending, third-point loading (57) with an effective stressed volume of 1.93 mm^3. The Weibull modulus for PyC is $\sim$10 (57).

Fluidized-bed isotropic PyCs are remarkably fatigue resistant. There is strong evidence for the existence of a fatigue threshold that is very nearly the single cycle

Table 3. Biomedical Fluidized-Bed Pyrolytic Carbon Properties

Property	Pure PyC	Typical Si-Alloyed PyC	Typical Glassy Carbon	POCO Graphite AXF-5Q
Flexural strength, MPa	493.7 ± 12	407.7 ± 14.1	175	90
Young's modulus, GPa	29.4 ± 0.4	30.5 ± 0.65	21	11
Strain-to-failure, %	1.58 ± 0.03	1.28 ± 0.03		0.95
Fracture toughness, $MPa \cdot \sqrt{m}$	1.68 ± 0.05	1.17 ± 0.17	0.5–0.7	1.5
Hardness, DPH, 500 g load	235.9 ± 3.3	287 ± 10	150	120
Density, $g \cdot cm^{-3}$	1.93 ± 0.01	2.12 ± 0.01	< 1.54	1.78
CTE, $\mu m \cdot cm^{-1}$ EC	6.5	6.1		7.9
Silicon content, %	0	6.58 ± 0.32	0	0
Wear resistance	Excellent	Excellent	Poor	Poor

fracture strength (58–60). Paris-law fatigue crack propagation rate exponents are high; on the order of 80 and da/dN fatigue crack propagation testing displays clear evidence of a fatigue crack propagation threshold (58–63).

Crystallographic mechanisms for fatigue crack initiation and damage accumulation are not significant in the PyC at ambient temperatures (59,61). There have been no clear instances of fatigue failure in a clinical implant during the accumulated 30-year experience (64). Less than 60 out of >4 million implanted PyC components have fractured (65), and these were caused by damage from handling or cavitation (66–68).

The PyC wear resistance is excellent. Wear testing performed in the 1970s identified titanium alloy, cobalt chromium alloy, and PyC as low wear contact materials for use in contact with PyC (69,70). This study determined that wear in PyC occurred due to an abrasive mechanism and interpreted wear resistance as approximately proportional to the ratio $H^2/2E$, where H is the Brinell hardness number and E is Young's modulus. This criteria is related to the amount of elastic energy that can be stored in the wearing surface (70). The greater the amount of stored energy, the greater the wear resistance. Successful low wearing contact couples used for mechanical heart valves include PyC against itself, cobalt chromium alloy, and ELI titanium alloy.

Observed wear in retrieved PyC mechanical heart valve prosthesis implant components utilizing PyC coupled with cobalt chromium alloy is extremely low with PyC wear mark depths of <2 μm at durations of 17 years (71–73). Wear in the cobalt chromium components was higher, 19 μm at 12 years (71–73). But, wear in the cobalt chromium components was concentrated at fixed contact points instead of being distributed over a large area as for the PyC rotating disk. Wear depths in all PyC prostheses, with fixed contact points are also low, <3.5 μm at 13 years (74,75). In contrast, the wear depths in valves utilizing polymeric components such as the polyacetyl Delrin in contact with cobalt chromium and titanium alloys are much higher at 267 μm at 17 years (76). Incorporation of PyC in heart valve prostheses has eliminated wear as a failure mode (29,77).

The PyC is often used in contact with metals and behaves as a noble metal in the galvanic series. Testing using mixed potential corrosion theory and potentiostatic polarization has determined that no detrimental effects occur for PyC coupled with titanium or cobalt–chrome alloys (78,79). Use of PyC with stainless steel alloys is not recommended.

To date, PyC has been used in ~ 25 mechanical heart valve prosthesis designs. One such design, the On-X valve, by Medical Carbon Research Institute, http://www.mcritx.com, is shown in Fig. 6.

Pyrolytic carbon has a good potential for orthopedic applications because of advantages over metallic alloys and polymers (3,30,31):

1. A modulus of elasticity similar to bone to minimize stress shielding.
2. Excellent wear characteristics.
3. Excellent fatigue endurance.
4. Low coefficient of friction.
5. Excellent biocompatibility with bone and hard tissue.
6. Excellent biocompatibility with cartilage.
7. Fixation by direct bone apposition.

A brief comparison of PyC properties to those of conventional/orthopedic implant materials is given in Table 4. Pyrolytic carbon coatings for orthopedic implants can reduce wear, wear particle generation, osteolysis aseptic loosening, and thus extend implant useful lifetimes. Furthermore, good PyC compatibility with bone and the native joint capsule enables conservative hemiarthroplasty replacements as an alternative to total joint replacement.

Cook et al. (80) studied hemijoint implants with a PyC femoral head in the canine hip and observed a greater potential for acetabular cartilage survival in PyC than for cobalt–chromium–molybdenum alloy and titanium alloy femoral heads. There were significantly lower levels of gross acetabular wear, fibrillation, eburnation, glycosaminoglycan loss, and subchondral bone change for PyC than the metallic alloys.

Tian et al. (81) surveyed *in vitro* and clinical *in vivo* PyC orthopedic implant studies conducted during the 1970s through the early 1990s and concluded that PyC demonstrated good biocompatibility and good function in clinical applications.

A 10-year follow-up of PyC metacarpophalangeal (MCP) finger joint replacements implanted in patients at the Mayo Clinic, Rochester Minnesota (82) between 1979 and 1987, demonstrated excellent performance. Ascension Orthopedics was able to use these results in part to justify a FDA premarket approval application (PMA) for the semiconstrained, uncemented MCP finger joint replacement, PMA P000057, Nov. 2001.

Figure 6. On-X prosthetic heart valves manufactured by Medical Carbon Research Institute from the elementally pure, fluidized-bed isotropic pyrolytic carbon, On-X carbon. The valves consist of a central flow circular orifice with two semicircular occluder disks. A polymeric sewing cuff is used to attach the valve to the annulus tissue. Aortic and mitral valves with two different sewing cuff designs each are shown.

Currently, Ascension Orthopedics, http://www.ascensionortho.com, manufactures PyC prostheses for finger joints: metacarpophalengeal (MCP) and proximal interphalangeal (PIP) in addition to carpometacarpal (CMC) thumb and an elbow radial head (RH) prostheses (see Fig. 7). Because of the excellent PyC compatibility with bone and cartilage, the CMC and radial head are used in hemiarthroplasty directly contacting the native joint capsule and bone. Fixation is by direct bone opposition for all of the prostheses. To date, $\sim$6500 Ascension Orthopedics prostheses have been implanted. Another company, Bioprofile, http://www.bio-profile.com, manufactures hemiarthroplasty PyC prostheses for the wrist: scapoid, scapho-trapezo-trapezoid, trapezium bone, capitate head, and an elbow radial head.

Glassy carbons have been proposed as an attractive low cost alternative for a variety of orthopedic and cardiovascular devices (3). However, because of relatively low strength and poor wear resistance it has not been generally accepted as a suitable material for long-term critical implants. An example of poor glassy carbon durability when used for heart valve components was cited earlier in the text (44).

Carbon fibers are popular as high strength fillers for polymers and other material composites and have been proposed for use in tendon and ligament replacements in addition to orthopedic and dental implants (83–86). Spinal interbody fusion cages using PEEK and carbon fibers (86) are an example of an orthopedic application. However, the ultimate properties of the implant depend largely upon the

Table 4. Material Properties of Orthopedic Materials

Property	Unit	PyC	Al_2O_3	TZP	CoCrMo	UHMWPE
Density	$g \cdot cm^{-3}$	1.93	3.98	6.05	8.52	0.95
Bend strength	MPa	494	595	1000	690, uts	20
Young's modulus, E	GPa	29.4	400	150	226	1.17
Hardness, H	HV	236[a]	2400	1200	496	NA
Fracture toughness, K_{1c}	$MN \cdot m^{-3/2}$	1.68	5	7		
Elongation at failure	%	2	0.15		1	>300
Poisson's ratio		0.28	0.2	0.2	0.3	
$H^2/2E^b$		7.6	12.2		1.8	

[a]The hardness value for PyC is a hybrid definition that represents the indentation length at a 500 g load with a diamond penetrant indentor. Because PyC elastically completely recovers the microhardness indentation a replica material such as a cellulose acetate coating, or a thin copper tape is used to "record" the fully recovered indentation length. Although unusual, this operational definition for hardness is a common practice used throughout the PyC heart valve industry.

[b]Approximate values, there are no exact conversions.

Figure 7. Ascension Orthopedics small joint PyC prostheses for finger joints.

matrix in which the carbon fibers are included and the geometry and orientation of the fiber inclusions (3).

Diamond-like carbon (DLC) coatings may find use as low friction, wear resistant surfaces for joint articulating surfaces in orthopedic implants (87,88). However, the coating thickness is limited to the micrometer level; the technology is still in development and ultimately may not be competitive with the newer ceramic joint replacement materials.

Buckyballs (fullerenes) and carbon nanotubes are cage-like structures that suggest use as a means to encapsulate and selectively deliver molecules to tissues. Because of their nanometer dimensions, fullerenes can potentially travel throughout the body. Some current biomedical applications under study involve functionalizing fullerenes with a number of substances including DNA and peptides that can specifically target, mark, or interfere with active sites on enzymes and perhaps inhibit virulent organisms such as the human immunodeficiency virus (89–93). They may also be used to selectively block ion channels on membranes (94). Fullerenes are synthesized by CVD and PVD techniques and can have a variety of novel properties depending on preparation. Currently, there are production difficulties with separation and isolation of fullerenes from the rest of soot-like materials that can occur during synthesis. However, bulk separation methods have been developed and some commercial sources have appeared. See http://www. chemistry.wustl.edu/~edudev/Fullerene/fullerene.html#index and http://www.pa.msu.edu/cmp/csc/nanotube. html. There is a wealth of information available on

the Internet that is readily accessed. Medical applications of fullerenes are currently a topic of intense interest and activity and hold much promise for future developments.

CONCLUSION

Uses of carbon as a biomaterial range from burnt toast, as mother's first aid remedy for suspected poisoning, to the newly discovered fullerene nanomaterials as a possible means to treat disease on a molecular level. The most successful and widespread medical applications have been the use of activated carbons for detoxification and the use of the General Atomics family of isotropic, fluidized-bed, pyrolytic carbons for structural components of long-term critical implants. However, the successful biomedical application of carbon requires an understanding that carbon is a spectrum of materials with wide variations in structure and properties. While a given carbon may be biocompatible, it may not have the mechanical and physical properties needed for the intended application.

As for the future, additional applications of the biomedical PyC materials to orthopedic applications in larger joints and in the spine can be expected, especially if successful long-term hemiarthroplasty devices can be demonstrated. New cardiovascular devices can be expected, such as components for venous shunts and venous valves. The most exciting new developments will probably occur in nanotechnology with the creation of functional, fullerene type, materials, devices, and systems through control of matter at the scale of 1–100 nm, and the exploitation of novel properties and phenomena at the same scale.

BIBLIOGRAPHY

Cited References

1. Pauling L. College Chemistry. San Francisco: W.H. Freeman; 1964.
2. Pierson HO. Handbook of Carbon, Graphite, Diamond and Fullerenes. Park Ridge, New Jersey: Noyes Publications; 1993.
3. Haubold AD, More RB, Bokros JC. Carbons. In: Black J, Hastings G, editors. Handbook of Biomaterial Properties. London: Chapman & Hall; 1998. p 464–477.
4. Janvier G, Baquey C, Roth C, Benillan N, Belisle S, Hardy J. Extracorporeal circulation, hemocompatibility, and biomaterials. Ann Thorac Surg 1996;62:1926–1934.
5. Morrison RT, Boyd RN. Organic Chemistry. Boston: Allyn and Bacon; 1974.
6. Properties and Characteristics of Graphite for the Semiconductor Industry. In: Sheppard RG, Mathes DM, Bray DJ, editors. Decatur, TX: POCO Graphite, Inc.; November 2001. Can be downloaded from http://www.poco.com.
7. Spain IL. Electronic Transport Properties of Graphite, Carbons, and Related Materials. Chem Phys Carbon 1981; 16:119
8. Kroto HW, Heath JR, O'Brien SC, Curl RF, Smalley RE. C_{60}: Buckminsterfullerene. Nature (London) 1985;318(6042): 162–163.
9. Haddon RC, Brus LE, Raghavachari K. Electronic Structure and Bonding in Icosahedral C_{60}. Chem Phys Lett 1986; 125:459.
10. O'Brien SC, Heath JR, Curl RF, Smalley RE. Photophysics of Buckminsterfullerene and Other Carbon Cluster Ions. J Chem Phys 1988;88:220.
11. Heath JR, O'Brien SC, Zhang Q, Liu Y, Curl RF, Kroto HW, Tittel FK, Smalley RE. Lanthanum Complexes of Spheroidal Carbon Shells. J Am Chem Soc 1985;107:7779–7780.
12. Chai Y, Guo T, Jin C, Haufler RE, Chibante LPF, Fure J, Wang L, Alford JM, Smalley RE. Fullerenes with Metals Inside. J Phys Chem 1991;95:7564
13. Heiney PA, Fischer JE, McGhie AR, Romanow WJ, Denenstein AM, McCauley JP, Jr., Smith AB, III, Cox DE. Orientational Ordering Transition in Solid C_{60}. Phys Rev Lett 1991;66:2911.
14. Haddon RC, Hebard AF, Rosseinsky MJ, Murphy DW, Duclos SJ, Lyons KB, Miller B, Rosamilia JM, Fleming RM, Kortan AR, Glarum SH, Makhija AV, Muller AJ, Eick RH, Zahurak SM, Tycko R, Dabbagh G, Thiel FA. Conducting Films of C_{60} and C_{70} by Alkali-Metal Doping. Nature (London) 1991; 350:320.
15. Bleeke JR, Frey RF. Fullerene Science Module. St. Louis, MO: Department of Chemistry, Washington University; Available at http://www.chemistry.wustl.edu~/edudev/Fullerene/fullerene.html.
16. Iijima S. Helical microtubules of graphitic carbon. Nature (London) 1991;354:56.
17. Tomanek D, of the University of Michigan, nanotube web page http://www.pa.msu.edu/cmp/csc/nanotube.html.
18. Jenkins GM, Kawamura K. Polymeric Carbons–Carbon Fibers, Glass and Char. Cambridge: Cambridge University Press; 1976.
19. Bokros JC. Deposition, Structure and Properties of Pyrolytic Carbon. In: Walker PL, editor. Chemistry and Physics of Carbon. Volume 5, New York: Marcel Dekker, Inc.; 1969. p 1–118.
20. Ely JL, Emken MR, Accuntius JA, Wilde DS, Haubold AD, More RB, Bokros JC. Pure Pyrolytic Carbon: Preparation and Properties of a New Material, On-X Carbon for Mechanical Heart Valve Prostheses. J Heart Valve Dis 1998;7:626–632.
21. Cooney DO. Activated Charcoal: Antidotal and Other Medical Uses. New York: Marcel Dekker; 1980.
22. Baker FS, Miller CE, Repik AJ, Tolles ED. Activated Carbon, in Kirk–Othmer. Encyc Chem Technol 1992;4:1015–1037.
23. Puri Balwant Rai. Chemsiorbed oxygen evolved as carbon dioxide and its influence on surface reactivity of carbons. Carbon 1966;4:391–400.
24. Cheremishoff NP, Moressi AC. Carbon adsorption applications. In: Cheremisinoff NP, Ellerbusch F, editors. Carbon Adsorption Handbook. Ann Arbor: Ann Arbor Science; 1978.
25. Pradhan BK, Sandle NK. Effect of different oxidizing agent treatments on the surface properties of activated carbons. Carbon 1999;37:1323–1332.
26. McQreey RL. Carbon electrodes: structural effects on electron transport kinetics. In: Bard AJ, editor. Electroanalytical Chemistry. New York: Dekker; 1991.
27. See Matweb.com for a variety of properties for engineering materials.
28. Diefendorf RJ, Stover ER. Pyrolytic Graphites. . .How structure affects properties. Metals Prog 1962;8 (May): 103–108.
29. Schoen FJ. Carbons in Heart Valve Prostheses: Foundations and clinical Performance. In: Zycher M, editor. Biocompatible Polymers, Metals and Composites. Lancaster PA: Technomic; 1983. p 240–261.
30. Bokros J. Carbon biomedical devices. Carbon 1977;15:355–371.
31. Haubold AD, Shim HS, Bokros JC. Carbon in Medical Devices. In: Williams DF, editor. Biocompatibility of Clinical Implant Materials. Volume 2, Boca Raton, FL: CRC Press; 1981. p 3–42.

32. More RB, Haubold AD, Bokros JC. Pyrolytic Carbon for Long-Term Medical Implants. In: Ratner B, Hoffman A, Schoen F, Lemons J, editors. Biomaterials Science: An Introduction to Materials in Medicine. 2nd ed. Academic Press; 2004.
33. More RB, Sines G, Ma L, Bokros JC. Pyrolytic Carbon. Encyclopedia of Biomaterials and Biomedical Engineering. Marcel Dekker; 2004.
34. Baier RE, Gott VL, Feruse A. Surface Chemical Evaluation of Thromboresistant Materials Before and After Venous Implantation. Trans Am Soc Artif Intern Organs 1970; 16:50–57.
35. Lee RG, Kim SW. Adsorption of Proteins onto Hydrophobic Polymer Surfaces: Adsorption Isotherms and Kinetics. J Biomed Mater Res 1974;8:251.
36. Nyilas E, Chiu TH. Artificial Surface/Sorbed Protein Structure/Hemocompatibility Correlations. Artif Organs 1978;2 (Suppl): 56–62.
37. Salzman EW, Lindon J, Baier D, Merril EW. Surface-Induced Platelet Adhesion, Aggregation and Release. Ann NY Acad Sci 1977;283:114.
38. Feng L, Andrade JD. Protein Adsorption on Low-Temperature Isotropic Carbon: I Protein Conformational Change Probed by Differential Sacnning Calorimetry. J Biomed Mater Res 1994;28:735–743.
39. Chinn JA, Phillips RE, Lew KR, Horbett Fibrinogen and Albumin Adsorption to Pyrolite Carbon. Trans Soc Biomater 1994;17:250.
40. Guglielmotti MB, Renou S, Cabrini RL. A histomorphometric study of tissue interface by laminar implant test in rats. Int J Oral Maxillofac Implants 1999;14:565–570.
41. Santavirta S, Takagi M, Gomez-Barrena E, Nevalainen J, Lassus J, Salo J, Konttinen YT. Studies of host response to orthopedic implants and biomaterials. J Long Term Eff Med Implants 1999;9:67–76.
42. Hill JB, Horres CR. The BD Hemodetoxifier: Particulate release and its significance. In: Chang TMS, editor. Artificial Kidney, Artificial Liver and Artificial Cells. New York: Plenum Press; 1978. p 199–207.
43. Williams DF. The Williams' Dictionary of Biomaterials. United Kingdom: Liverpool University Press; 1999.
44. Fettel BE, Johnston DR, Morris PE. Accelerated life testing of prosthetic heart valves. Med Inst 1980;14(3): 161–164.
45. Bensen J. Pre-Survey on the Biomedical Applications of Carbon. 1969. North American Rockwell Corporation Report R-7855.
46. Ford X. Clinical Toxicology. 1st ed., W. B. Saunders Company; 2001.
47. Roberts X. Clinical Procedures in Emergency Medicine. 3rd ed., W. B. Saunders Company; 1998.
48. LaGrange LD, Gott VL, Bokros JC, Ramos MD. Compatibility of Carbon and Blood. In: Hegyeli RJ, editor. Artificial Heart Program Conference Proceedings. Washington, DC: US Government Printing Office; 1969. Chapt. 5. p 47–58.
49. Sadeghi H. Dysfonctions des prostheses valvulaires cardaques et leur traitment chirgical. Schwiez Med Wschr 1987; 117:1665–1670.
50. Kaae JL. The mechanism of deposition of pyrolytic carbon. Carbon 1985;23(6): 665–667.
51. Kaae JL, Wall DR. Microstructural Characterization of Pyrolytic Carbon for Heart Valves. Cells Mater 1996;6(4): 281–290.
52. Ma L, Sines G. High resolution structural studies of a pyrolytic carbon used in medical applications. Carbon 2002;40:451–454.
53. Akins RJ, Bokros JC. The Deposition of Pure and Alloyed Isotropic Carbons and Steady State Fluidized Beds. Carbon 1974;12:439–452.
54. Reilly DT, Burstein AH, Frankel VH. The Elastic Modulus for Bone. J Biomech 1974;7:271.
55. Reilly DT, Burstein AH. The Mechanical Properties of Bone. J Bone Jt Surg Am 1974;56:1001.
56. De Salvo G. Theory and Structural Design Applications of Weibull Statistics. 1970. WANL-TME-2688, Westinghouse Electric Corporation.
57. More RB, Kepner JL, Strzepa P. Hertzian Fracture in Pyrolite Carbon. In: Ducheyne P, Christiansen D, editors. Bioceramics. Volume 6, Oxford: Butterworth-Heinemann Ltd; 1993. p 225–228.
58. Gilpin CB, Haubold AD, Ely JL. Fatigue Crack Growth and Fracture of Pyrolytic Carbon Composites. In: Ducheyne P, Christiansen D, editors. Bioceramics. Volume 6, Oxford: Butterworth-Heinemann Ltd; 1993. p 217–223.
59. Ma L, Sines G. Fatigue of Isotropic Pyrolytic Carbon Used in Mechanical Heart Valves. J Heart Valve Dis 1996;5(Suppl.I): S59–S64.
60. Ma L, Sines G. Unalloyed Pyrolytic Carbon for Implanted Heart Valves. J Heart Valve Dis 1999;8(5): 578–585.
61. Ma L, Sines G. Fatigue Behavior of Pyrolytic Carbon. J Biomed Mat Res 2000;51:61–68.
62. Ritchie RO, Dauskardt RH, Yu W, Brendzel AM. Cyclic Fatigue-crack Propagation, Stress Corrosion and Fracture Toughness Behavior in Pyrolite Carbon Coated Graphite for Prosthetic Heart Valve Applications. J Biomed Mat Res 1990;24:189–206.
63. Beavan LA, James DW, Kepner JL. Evaluation of Fatigue in Pyrolite Carbon. In: Ducheyne P, Christiansen D, editors. Bioceramics. Volume 6, Oxford: Butterworth-Heinemann Ltd; 1993. p 205–210.
64. Bokros JC, Haubold AD, Akins RJ, Campbell LA, Griffin CD, Lane E. The durability of mechanical heart valves replacements: past experience and current trends. In: Bodnar E, Frater RWM, editors. Replacement Cardiac Valves. New York: Pergamon Press; 1991. p 21–47.
65. Haubold AD. On the Durability of Pyrolytic Carbon In Vivo. Med Prog Through Technol 1994;20:201–208.
66. Kelpetko V, Moritz A, Mlczoch J, Schurawitzki H, Domanig E, Wolner E. Leaflet Fracture in Edwards-Duromedics Bileaflet Valves. J Thorac Cardiovasc Surg 1989;97: 90–94.
67. Kafesjian R, Howanec M, Ward GD, Diep L, Wagstaff L, Rhee R. Cavitation Damage of Pyrolytic Carbon in Mechanical Heart Valves. J Heart Valve Dis 1994;3(Suppl I): S2–S7.
68. Richard G, Cao H. Structural failure of Pyrolytic Carbon Heart Valves. J Heart Valve Dis 1996;5(Suppl I): S79–S85.
69. Shim HS, Schoen FJ. The wear resistance of pure and silicon-alloyed isotropic carbons. Biomater Med Dev Art Org 1974;2(2): 103–118.
70. Shim HS. The wear of titanium alloy, and UHMW polyethylene caused by LTI carbon and Stellite 21. J Bioengr 1977;1:223–229.
71. More RB, Silver MD. Pyrolytic Carbon Prosthetic Heart Valve Occluder Wear: In Vivo vs. In Vitro Results for the Björk-Shiley Prosthesis. J Appl Biomater 1990;1:267–278.
72. More RB. An Examination of Two Retrieved Long-Term Human Implant Bjork-Shiley Valves. Med Prog Technol 1994;20:195–200.
73. More RB, Haubold AD, Silver MD. Pyrolytic Carbon Wear in Retrieved Mechanical Heart Valve Prosthesis Implants. 25th Annual Meeting of the Society for Biomaterials, 1999. p 553.
74. More RB, Chang BC, Hong YS, Cao BK, Butany J, Wear Analysis of Retrieved Mitral Bileaflet Mechanical Heart Valve Prostheses, Presented to the Society for Heart Valve Disease, 1st Biennial Symposium, London; June 2001.
75. More RB, Haubold AD, Silver MD. Pyrolytic Carbon Wear in Retrieved Mechanical Heart Valve Prosthesis Implants. 25th Annual Meeting of the Society for Biomaterials, 1999. p 553.

76. Wieting DW. The Björk-Shiley Delrin Tilting Disc Heart Valve: Historical Perspective, Design and Need for Scientific Analyses After 25 Years. J Heart Valve Dis 1996;5(Suppl I): S157–S168.
77. Schoen FJ, Titus JL, Lawrie GM. Durability of Pyrolytic Carbon-Containing Heart Valve Prostheses. J Biomed Mater Res 1982;16:559–570.
78. Griffin CD, Buchanan RA, Lemons JE. In Vitro Electrochemical Corrosion Study of Coupled Surgical Implant Materials. J Biomed Mater Res 1983;17:489–500.
79. Thompson NG, Buchanan RA, Lemons JE. In Vitro Corrosion of Ti-6Al-4V and Type 316L Stainless steel When Galvanically Coupled with Carbon. J Biomed Mater Res 1979;13:35–44.
80. Cook SD, Thomas KA, Kester MA. Wear characteristics of the canine acetabulum against different femoral prostheses. J Bone Joint Surg 1989;71B:189–197.
81. Tian CL, Hetherington VJ, Reed S. A Review of Pyrolytic carbon: Application in Bone and Joint Surgery. J Foot Ankle Surg 1993;32(5):490–498.
82. Cook SD, Beckenbaugh RD, Redondo J, Popich LS, Klawitter JJ, Linscheid RL. Long term follow-up of pyrolytic carbon metacarpophalangeal implants. J Bone Joint Surg 1999; 81A(5): 635–648.
83. Ferrari M. Clinical evaluation of fiber-reinforced epoxy resin posts and cast post and cores. Am J Dent 2000; 01-May-13 (Spec No): 15B–18B.
84. Pamula E. Studies on development of composite biomaterials for reconstruction of the larynx. Polim Med 2001;31(1–2):39–44.
85. Katoozian H. Material optimization of femoral component of total hip prosthesis using fiber reinforced polymeric composites. Med Eng Phys 2001;23(7):503–509.
86. Früh HJ. Fusion implants of carbon fiber reinforced plastic. Orthopade 2002;31(5):454–458.
87. Dearnaley G. Diamond-like carbon: a potential means of reducing wear in total joint replacements. Clin Mater 1993;12:237–244.
88. Lappalainen R, Anttila A, Heinonen H. Diamond coated total hip replacements. Clin Orthop 352 (July 1998): 118–127.
89. Pantarotto D, Partidos CD, Graff R, Hoebeke J, Briand JP, Prato M. Bianco A. Synthesis, structural characterization, and immunological properties of carbon nanotubes functionalized with peptides. J Am Chem Soc 2003 May 21; 125(20): 6160–6164.
90. Qingnuan L, Yan X, Xiaodong Z, Ruili L, Qieqie D, Xiaoguang S, Shaoliang C, Wenxin L. Preparation of (99m)Tc-C(60)(OH)(x) and its biodistribution studies. Nucl Med Biol 2002 Aug; 29(6): 707–710.
91. Gonzalez KA, Wilson LJ, Wu W, Nancollas GH. Synthesis and in vitro characterization of a tissue-selective fullerene: vectoring C(60)(OH)(16)AMBP to mineralized bone. Bioorg Med Chem 2002 Jun; 10(6): 1991–1997.
92. Wolff DJ, Barbieri CM, Richardson CF, Schuster DI, Wilson SR. Trisamine C(60)-fullerene adducts inhibit neuronal nitric oxide synthase by acting as highly potent calmodulin antagonists. Arch Biochem Biophys 2002 Mar 15; 399(2): 130–141.
93. Schinazi RF, Sijbesma R, Srdanov G, Hill CL, Wudl F. Synthesis and virucidal activity of a water-soluble, configurationally stable, derivatized C60 fullerene. Antimicrob Agents Chemother 1993 Aug.; 37(8): 1707–1710.
94. Park KH, Chhowalla M, Iqbal Z, Sesti F. Single-walled carbon nanotubes are a new class of ion channel blockers. J Biol Chem 2003 Dec. 12; 278(50): 50212–50216, Epub 2003 Sep. 30.

See also BIOMATERIALS: TISSUE ENGINEERING AND SCAFFOLDS; BIOMATERIALS, TESTING AND STRUCTURAL PROPERTIES OF; BIOSURFACE ENGINEERING; MATERIALS AND DESIGN FOR ORTHOPEDIC DEVICES; HEART VALVE PROSTHESES.

BIOMATERIALS CORROSION AND WEAR OF

ROGER J. NARAYAN
University of North Carolina
Chapel Hill, North Carolina

MIROSLAV MAREK
Georgia Institute of Technology
Atlanta, Georgia

CHUNMING JIN
North Carolina State University
Raleigh, North Carolina

INTRODUCTION

Many materials suffer degradation with time when exposed to aggressive chemical environments within the human body. In metallic biomaterials, degradation results from electrochemical corrosion. Ceramic and polymeric biomaterials may undergo physical or chemical deterioration processes. In addition, mechanical forces may act to increase damage by wear, abrasion, or environment-induced cracking processes.

Corrosion of implants, dental restorations, and other objects placed in the human body may result in degradation of function as a result of loss of mass, decrease in mechanical integrity, or deterioration of aesthetic qualities. The associated release of corrosion products and the flow of the corrosion currents also may cause inflammation, allergic reactions, local necrosis, and many other health problems.

For electronic conductors (e.g., metals), corrosive interaction with ionically conducting liquids (e.g, body fluids) is almost always electrochemical. The degradation of metals is due to an oxidation process that involves the loss of electrons. This process involves a change from a metallic state to an ionic state, in which the ions dissolve or form nonmetallic solid products. For the process to continue, the released electrons must be consumed in a complementary reduction, which usually involves species present in the biological environment (e.g., hydrogen ions or dissolved oxygen). The reaction resulting in oxidation is usually called an anodic process and reaction resulting in reduction is usually called a cathodic process. The metal is referred to as an electrode, and the liquid environment is referred to as an electrolyte.

For many metals, the most important environmental variables are the concentrations of chloride ions, hydrogen ions, and dissolved oxygen. In many human body fluids, the chloride ion concentration varies in a relatively narrow range near $0.1\,\mathrm{mol \cdot L^{-1}}$; however, it may be variable (e.g., urine) or lower (e.g., saliva) in certain body fluids. The hydrogen ion concentration is expressed as a pH value and is near neutral ($\mathrm{pH} = 7$) for plasma, interstitial fluid, bile, and saliva; however, it is more variable ($\mathrm{pH} = 4$–8) in urine and very low ($\mathrm{pH} = 1$–3) in gastric juice (1). The chloride concentration and pH are most important factors determining the rate of oxidation because of their effect on protective oxide passivating films on metals. The dissolved oxygen concentration affects mainly the cathodic process. The usual range of partial pressure of oxygen in body fluids is ~40–100 mmHg (5.33–13.33 kPa) (1–3).

For electrochemical oxidation to cause clinically relevant degradation of a material, the electrochemical reaction must be energetically possible (thermodynamics) and the reaction rate must be appreciable (kinetics). Oxidation of nickel, for example,

$$Ni \rightarrow Ni^{2+} + 2e^{-} \quad (1)$$

will proceed in the indicated direction if the potential of the electrode on which the reaction occurs is higher (more positive or less negative) than the equilibrium potential for a given electrolyte, and is a function of the energy change involved. The equilibrium potential also depends on temperature, pressure, and activity ($\approx$ concentration) of ions. The values of potentials for reactions between metals and their ions in water are given under standard conditions (temperature 25 °C, pressure 1 atm, and activity of ions equal to 1) and are written in the form of materials reduction. These values, known as standard single electrode potentials, are listed in the so-called electrochemical or electromotive (EM) series (1). Noble metals, which have no tendency to dissolve in water have positive standard single-electrode potentials. On the other hand, active metals with a high tendency to react with water exhibit negative potentials.

While the potential of the anodic process must be above the equilibrium potential for the reaction to proceed as oxidation, for the cathodic process the electrode potential must be below (more negative or less positive) the equilibrium potential for net reduction to occur. For reduction of hydrogen ions,

$$2H^{+} + 2e^{-} = H_2 \quad (2)$$

the equilibrium potential at normal body temperature (37 °C) and pH 7.4 (blood or interstitial fluids) is −0.455 V (SHE) (−0.697 V, SCE), while the equilibrium potential of the other likely cathodic reaction,

$$O_2 + 4\,H^{+} + 4e^{-} = 2\,H_2O \quad (3)$$

is −0.753 V (SHE) (0.511 V, SCE) at 40 mmHg (5.33 kPa) of oxygen partial pressure. Although reaction 3 is more sluggish than reaction 2, for most metals in the human body the electrode potential of reaction 3 is above the equilibrium potential of the hydrogen reaction 2, and reduction of oxygen is the dominant cathodic process.

In spontaneous electrochemical corrosion, at least two reactions occur simultaneously. At least one reaction occurs in the direction of oxidation, and at least one reaction occurs in the direction of reduction. Each reaction has its own equilibrium potential, and this potential difference results in a current flow, as the electrons released in oxidation flow to the sites of reduction and are consumed there. In the absence of a significant electrical resistance in the current path between the reaction sites, a common potential is established, which is known as mixed potential or corrosion potential (E_{corr}). At this potential, both reactions produce the same current in opposite directions in order to preserve electrical neutrality. The value of the oxidation current, which is equal to the absolute value of the reduction current, per unit area at this potential is known as the corrosion current density (i_{corr}). The oxidation and reduction reactions may be distributed uniformly on the same metal surface; however, there are often some regions of the biomaterial surface that are more favorable for oxidation and other regions that are more favorable for reduction. As a result, either local anodic and cathodic areas or completely separate anodes and cathodes are formed.

The corrosion rate (mass of metal oxidized per unit area and time) is proportional to the corrosion current density. The conversion is given by the Faraday's law, which states that an electric charge of 96,485 C is required to convert 1 equiv weight of the metal into ions, or vice versa. The shift of the potential of a reaction from the equilibrium value to the corrosion potential is called polarization by the flow of the current. The resulting current flowing at corrosion potential depends on the way the current of each reaction varies with the potential. If the current is controlled by the activation energy barrier for the reaction at the electrode surface, then the reaction rate increases exponentially with increasing potential for oxidation reactions and decreases exponentially with increasing potential for reduction reactions. The activation energy controlled current typically increases or decreases ten times for a potential change of ~50–150 mV. At high reaction rates, the current may become limited by the transport of reaction species to or from the electrodes; eventually, the corrosion process may become completely controlled by diffusion and independent of potential.

The vast majority of uses for metallic biomaterials in the human body are successful due to the phenomenon of passivity. In a passive state, these metals become covered with thin, protective films of stable, poorly soluble oxides or hydroxides when exposed to an aqueous electrolyte. Once this passivating film forms, the current density drops to a very low value and becomes much less dependent on the potential. The variation of the reaction current density with the potential can be illustrated in a polarization diagram. A schematic diagram in Fig. 1 shows some of the main reactions in corrosion and relevant parameters. A straight-line relationship in a semilogarithmic (E vs. log I) diagram indicates that an activation energy-controlled reaction is occurring. This electrochemical activity is known as Tafel behavior, and the slopes of the lines (~50–150 mV per 10-fold change in the current or current density) are equal to the values of the Tafel constants. When the oxidation reaction of the metal shows this relationship at the corrosion potential, it indicates that the metal is actively corroding. If a metal forms a passivating film when the potential exceeds a critical value in the active corrosion region, then the current density drops from a value called the critical current density for passivation (i_{crp}) at a primary passivation potential (E_{pp}) to a low current density in the passive state (i_p). This behavior is illustrated schematically in Fig. 2. For an electrode to maintain a stable passive state, the intersection of the oxidation (anodic) and reduction (cathodic) lines must occur in the region of passivity.

The polarization characteristics of a biomaterial can be experimentally determined using a device called a potentiostat, which maintains the sample potential at a set value

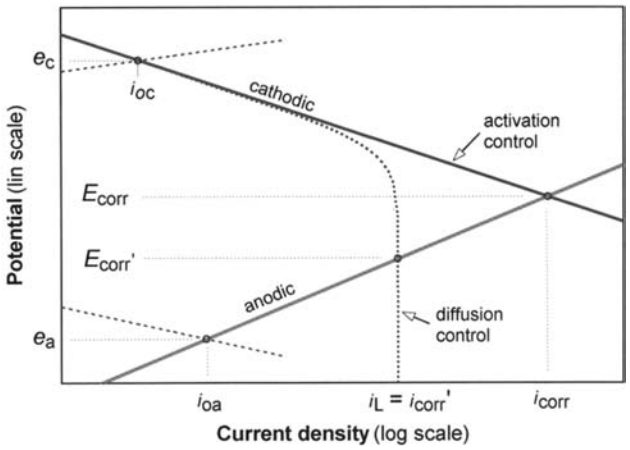


Figure 1. Schematic polarization diagram showing oxidation (anodic) and reduction (cathodic) reactions of a corrosion process, for reactions controlled by activation energy and by mass transport (diffusion). In this figure, e_a and e_c refer to equilibrium potentials of the anodic and cathodic process, respectively, i_{oa} and i_{oc} refer to exchange current densities, E_{corr} refers to mixed corrosion potential, and i_L refers to limiting current density.

versus a reference electrode by passing current between the sample and an auxiliary counterelectrode. A scan generator can be used to vary the controlled potential over a range of interest, and the E–i relationship can be determined. The relationship of main interest is usually the oxidation rate as a function of the potential, which can be depicted in an anodic polarization diagram. Since only a net current (difference between the absolute values of the oxidation and reduction currents) can be measured, the experimental polarization curve shows a value approaching zero at the intersection of the anodic and cathodic polarization curves.

Experimental anodic polarization curves for passivating metals and alloys often do not exhibit the passivation peak shown in Fig. 2, either because the metal forms an oxide in the electrolyte without undergoing active dissolution or because an oxide film already has formed as a result of exposure to air. More importantly for human body fluids and other chloride-containing electrolytes, the region of passivity is often limited by a localized passivation breakdown above a critical breakdown potential (E_b). When a breakdown occurs, intensive oxidation takes place within localized regions on the biomaterial surface, resulting in sometimes significant pit formation. In an experimental anodic polarization diagram, breakdown appears as a sharp increase in the measured current above the critical breakdown potential. Because of the destructive nature of surface pitting, the determination of critical breakdown potential is one of the most important ways of assessing the suitability of novel metallic biomaterials for use in medical devices.

The high current density in active pits is due to the absence of a passivating film, which results from local chemical and electrochemical reactions that change the electrolyte to become highly acidic and depleted in dissolved oxygen. A similar corrosion mechanism may occur in interstices known as crevices, where the transport of species to and from the localized corrosion cell is difficult. This process, known as crevice corrosion, does not require a potential exceeding the critical breakdown potential for the initiation of corrosion. Both pit and crevice corrosion cells may repassivate if the potential is lowered below a value needed for maintenance of a high oxidation rate on the bare (nonpassivated) metal surface. The potential below which active pits repassivate is called the repassivation or protection potential (E_p). The concept of a protection potential also applies to crevice corrosion. Experimentally, repassivation can be studied by reversing the anodic polarization scan and recording the potential at which the current returns to a passive state value (Fig. 3). Repassivation can also be examined by initiating pitting or crevice corrosion and lowering the potential in steps until the current

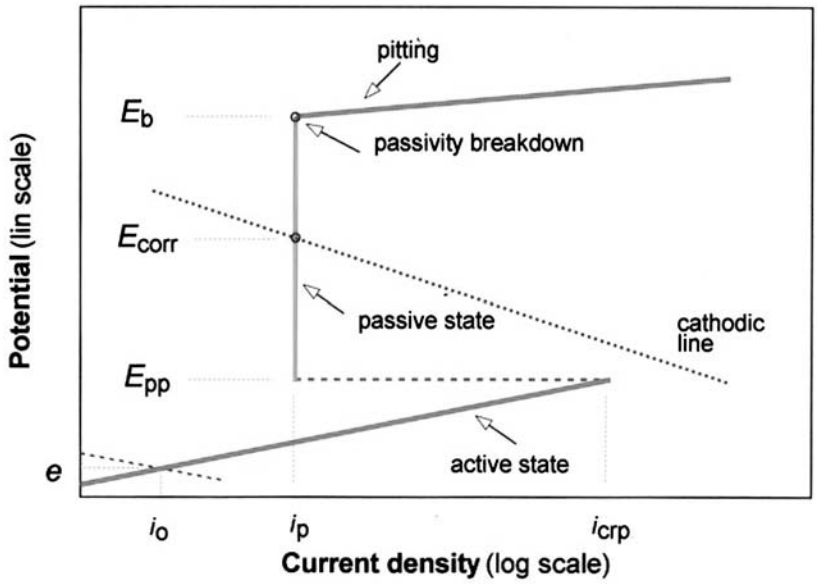


Figure 2. Schematic polarization diagram, showing a transition from active to passive state and a breakdown of passivity. In this figure, i_{crp} refers to critical current density for passivation, i_p refers to current density in the passive state, E_{pp} refers to primary passivation potential, and E_b refers to breakdown potential.

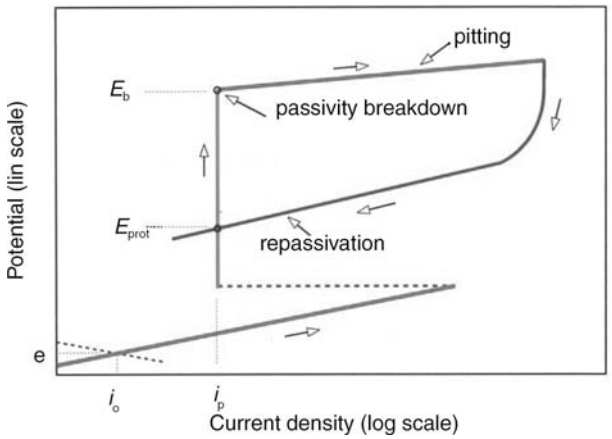


Figure 3. Schematic experimental cyclic polarization diagram for a passivating electrode, showing passivity breakdown and repassivation after potential scan reversal. In this figure, E_{prot} refers to protection (repassivation) potential.

shows low values that decrease with time (standard test methods F2129 and F746, respectively, ASTM 2005) (4). The difficulty in finding a reliable protection potential value is due to the fact that the ease of repassivation depends on the extent of pitting or crevice corrosion damage that has occurred before the potential drop.

For some polyvalent metals (e.g., chromium), soluble species (e.g., CrO_4^{2-}) become thermodynamically stable as the valence changes (e.g., from 3 to 6 in chromium) at potentials above those for a stable oxide. This process may result in another region of active dissolution at high potentials in a phenomenon known as transpassivity.

When corrosion is relatively uniform throughout the biomaterial, the most important corrosion parameter is the average corrosion rate. For biomaterials with very low corrosion rates, the average corrosion rate is mostly determined by sensitive electrochemical techniques. The average corrosion current density (i_{corr}) is usually determined either from the results of the polarization scan by extrapolating the anodic and cathodic lines to the corrosion potential or from calculating the value of the polarization resistance. The polarization resistance (R_p) is defined as the slope of the polarization curve at zero current density $[R_p = (dE/di_n)_{i_n=0}]$, in which i_n is the net (measured) current density.

When two or more dissimilar metals are placed in contact within an electrolyte, their interaction may cause galvanic corrosion. The oxidation degradation is enhanced for the metal with the lower individual corrosion potential, which becomes the anode of the cell. It is polarized towards a higher potential at the other electrode, which becomes the cathode. Since the oxidation current increase on the anode must be balanced by an identical reduction current increase on the cathode, a combination of a small anode with a large cathode is more detrimental than the reverse situation, since a larger increase in the oxidation current density is produced. In practical situations, resistance in the current path between the electrodes often reduces the galvanic effect. Differences in the concentrations of reaction species at different regions on the metal surface may result in a potential difference, which leads to additional polarization. An increase in the oxidation current density, a difference in the equilibrium potentials, and a flow of current may result. Differences in concentrations of hydrogen ions, dissolved metal ions, or dissolved oxygen may result in concentration cell corrosion.

Metal parts subjected to mechanical loading in a corrosive environment may fail by environment-induced cracking (EIC). Stress corrosion cracking (SCC) may occur in some biomaterials when they are subjected to static loading under certain environmental conditions. Corrosion fatigue (CF) may result from variable loading in reactive environments. When the failure can be attributed to the entry of hydrogen atoms into the metal, the phenomenon is referred to as hydrogen induced cracking (HIC). Environment-induced cracking may be caused by complex combinations of mechanical, chemical, and electrochemical forces; however, the exact mechanisms of this behavior are subject to significant controversy. In these cases, mechanical factors may play important roles in crack propogation.

Intergranular corrosion occurs when dissolution is confined to a narrow region along the grain boundaries. This process is either due to precipitation of corrosion susceptible phases or due to depletion in elements that provide corrosion protection along the boundaries, which is caused by precipitation of phases rich in those elements. Some stainless steel and nickel-chromium alloys may be sensitized due to precipitation of chromium-rich carbides along grain boundaries when heated to a specific temperature range. Sensitization is normally prevented from occurring in stainless steels currently used in medical devices, which contain very low amounts of carbon.

Passivating films may also be mechanically destroyed in wear- , abrasion- , erosion- , and fretting-corrosion processes. Wear-corrosion involves materials in a friction contact that exhibit substantial relative movement. Fretting occurs in situations in which there are only small relative movements between materials that are essentially fixed with respect to one another. The resulting wear debris may cause abrasion–corrosion behavior. Wear-corrosion may occur in artificial joints, including the metal ball of a hip joint in contact with the polyethylene cup. Fretting may take place between the ball and the stem of multicomponent hip implants. In both forms of corrosion, the narrow gap between contacting surfaces creates crevice conditions. In addition, the destructive effect of friction and abrasion on the protective surface film is superimposed on the corrosion mechanism in the crevice cell. Erosion corrosion may occur on devices exposed to rapidly flowing fluids, including the surfaces of artificial heart valves.

A wide variety of metals and alloys have been used in medical devices. The three most commonly used alloys are stainless steel, cobalt alloys, and titanium alloys (5). Type 316 LVM (low carbon, vacuum-melted) stainless steel is less corrosion resistant than cobalt or titanium alloys, and it is most often used for temporary implants (5–8). This material is referred to as an austenitic steel, because it contains an iron carbide phase called austenite (γ-iron). Implant-grade steel has a nominal composition of 18% chromium, 14% nickel, and 2.5% molybdenum; the

compositional limits and properties are specified by ASTM standards F 138 and F 139 for wrought steel and F 745 for cast steel (ASTM, 2005) (4). Chromium serves to improve corrosion resistance through the formation of a highly protective surface film rich in chromium oxide. Implant-grade steel has a low carbon content in order to prevent sensitization and intergranular corrosion. Alloying with - molybdenum further improves the resistance, especially to crevice corrosion and pitting. Nickel serves to stabilize the face-centered cubic (fcc) structure. On the other hand, manganese sulfide inclusions, which contribute to initiation of pitting, are minimized.

The corrosion resistance of stainless steel greatly depends on the surface conditions, and stainless steel implants are almost always electropolished and prepassivated by exposure to nitric acid (standard practice F86, ASTM 2005) (4). The breakdown potential is usually around 0.4 V (SCE), with a large hysteresis loop and a low protection potential (9). Considering that the potential in the human body is not likely to exceed about 0.5 V (SCE) (see Eq. 3 and its equilibrium potential), a well polished and passivated 361 LVM stainless steel is not very susceptible to pitting in the human body, especially for unshielded and undisturbed implant surfaces. Once localized attack is initiated, however, repassivation is difficult. As a result, stainless steel implants are very susceptible to crevice corrosion, especially when the crevice situation is combined with destruction of the surface film (e.g., fretting of bone plates under the screw heads). Small single component stainless steel implants, such as balloon-expandable vascular stents, that are made of high purity precursor materials and are subjected to a high quality surface treatment and inspection can achieve a breakdown potential in excess of 0.8 V (SCE); these materials are considered very resistant to localized corrosion (10). Stainless steel bars [containing 22% chromium 12.5% nickel , 5% manganese, and 2.5% molybdenum (ASTM F 1586)] and wires [containing 22% chromium, 12.5% nickel, 5% manganese, and 2.5% molybdenum (ASTM F 1314)] strengthened with nitrogen have shown a higher breakdown potential than ASTM F 138 steel (4).

Vitallium and other cobalt–chromium alloys were developed as a corrosion resistant, high strength alternative to stainless steel alloys. These materials were first used in dentistry, and were later introduced to orthopedics and other surgical specialties. The cast cobalt–chromium alloy most commonly used in medical devices (ASTM F 75) contains 28% chromium and 6% molybdenum (4). This alloy was found to be suitable for investment casting into intricate shapes. In addition, it exhibited very good corrosion and excellent wear resistance; however, it possessed low ductility. Alloys with slightly modified compositions were later developed for forgings (ASTM F 799) and wrought bars, rods, and wire (ASTM F 1537) (4). Alloy F75 has shown corrosion resistance superior to stainless steel in the human body. Laboratory studies reported a breakdown potential of 0.5 V (SCE) and protection potential of 0.4 V (SCE) (6,7,9,11). These properties have made it possible to use cobalt–chromium alloys for permanent implants. Cobalt–chromium alloys with porous surfaces have been used for bone ingrowth, although they have been superseded by even more crevice corrosion resistant and biocompatible titanium alloys. The excellent corrosion resistance of cobalt–chromium alloys can be attributed to a high chromium content and a protective surface film of chromium oxide. Concerns have been raised, however, regarding the release of biologically active hexavalent chromium ions (12). Other cobalt-based wrought surgical alloys include F90 (Co-Cr-W-Ni), F563 (Co-Ni-Cr-Mo-W-Fe), F563 (Co-Ni-Cr-Mo-W-Fe), F1058 (Co-Cr-Ni-Mo), and F688 (Co-Ni-Cr-Mo) (4). These alloys provide good to excellent corrosion behavior and a variety of mechanical properties, which depend on thermomechanical treatment. However, there is some concern regarding metal ion release in these alloys, which contain high nickel concentrations.

Titanium and titanium alloys have been used in orthopedic implants and other medical devices since the 1960s. Their popularity has rapidly increased because they possess high corrosion resistance, adequate mechanical properties, and relatively benign degradation products. Although titanium is thermodynamically one of the least stable structural metals in air and water, it acquires high resistance to corrosion due to a very protective titanium oxide film. Unalloyed titanium (ASTM F67 and F1341) and titanium-6 % aluminum, 4 % vanadium alloy (ASTM F136 and F1472 for wrought alloy and F1108 for castings) are commonly used in orthopedic prostheses (4). These materials exhibit a breakdown potential in body fluid substitutes well above the physiological range of potentials (several volts vs. SHE). In addition, they readily repassivate in biological fluids, which makes them highly resistant to pitting and crevice corrosion. The high crevice corrosion resistance and biocompatibility of titanium alloys have made it possible to create porous titanium surfaces that allow for bone ingrowth and cementless fixation of implants.

One shortcoming of titanium and titanium alloys is their relatively poor wear resistance (5). Since resistance to corrosion depends on the integrity of the protective oxide film, wear-corrosion remains a problem for titanium alloy prostheses. Surface treatments (including nitrogen diffusion hardening, nitrogen ion implantation, and thin-film deposition) may be used to provide more wear-resistant articulating surfaces. Another solution to titanium wear involves the use of multicomponent implants (e.g., implants that contain smooth surfaces made of cobalt–chromium alloy for articulating components and porous surfaces made out of titanium alloy for bone ingrowth and biological fixation). However, fretting corrosion may occur as a result of micromovement at the taper joints between the components, which may destroy the surface passivating films and increase overall corrosion rates (13–15). In spite of the very successful use of the Ti-6Al-4V alloy orthopedic implants, some concern remains regarding the possible toxicity of the aluminum and vanadium components within this alloy. A variety of vanadium-free or aluminum- , and vanadium-free alloys have been developed, including Ti-15Sn-4Nb-2Ta-0.2Pd, Ti-12Mo-6Zr-2Fe (TMZF), Ti-15Mo, and Ti-13Nb-13Zr (5). Ti-12Mo-6Zr-2Fe (TMZF) and Ti-13Nb-13Zr alloys exhibit lower elastic moduli and higher tensile properties. The alloying

elements also form highly protective oxides, which contribute to the excellent corrosion resistance of these materials (16).

An equiatomic nickel–titanium alloy (Nitinol) has received considerable interest as an implant material because of its shape memory and pseudoelasticity properties, the latter resulting in a very low apparent elastic modulus. This superelastic behavior has allowed the development of self-expandable vascular stents, bendable eyeglass frames, orthodontic dental archwire, and intracranial aneurysm clips. Several studies have shown good biocompatibility of Nitinol; however, clinical failures have also been reported (17–19). Laboratory studies have shown a wide variety of performance, with resistance to the breakdown of passivity ranging from poor to excellent (20–22). Resistance to the initiation of pitting critically depends on the surface conditions. A surface film that consists mostly of titanium oxide results in a high resistance to pitting; however, the presence of elemental nickel or nickel oxide reduces the breakdown potential. In addition, recent studies have shown that strained nickel–titanium alloy exhibits significant improved corrosion resistance over as-prepared materials. Other conditions that may affect corrosion resistance include surface roughness, the presence of inclusions, and the concentration of intermetallic species (23).

Another group of biomaterials is used in restorative dentistry and orthodontics. Materials for restorative dentistry must not only meet corrosion, wear, and compatibility considerations described earlier, but also satisfy aesthetic requirements and must have the capacity to be either precisely cast into intricate shapes or used to directly fill a prepared cavity in a tooth. Dental cast alloys can be roughly divided into three major groups of high noble alloys, seminoble alloys, and base alloys. The high noble alloys include those with a high percentage of gold or other noble metals (e.g., platinum), and derive their corrosion resistance mainly from a low thermodynamic tendency to react with the environment. Seminoble alloys often have complex compositions, and either possess a relatively low noble metal content or contain a significant concentration of silver. These materials possess a higher thermodynamic tendency to react than high noble alloys; however, their kinetics of aqueous corrosion in saliva is sufficiently slow, and allows these materials to provide adequate corrosion resistance under biological conditions. The main corrosion concern for seminoble alloys is their tendency to react with sulfur in food and drinks and form dark metallic sulfide film, resulting in the loss of aesthetic quality. Base dental cast alloys include cast titanium, titanium alloys, and nickel–chromium alloys. These materials lack the aesthetic qualities of noble alloys; however, they are resistant to sulfide tarnishing. Nickel–chromium alloys exhibit passivation behavior and some susceptibility to pitting and crevice corrosion. Cast titanium and titanium alloys exhibit highly protective passive films and high resistance to chloride corrosion; however, they demonstrate some susceptibility to fluoride attack, which is of some concern due to the prophylactic use of fluoride rinses and gels. Direct-filling metallic materials include unalloyed gold and dental amalgams, which are alloys of mercury, silver, tin, copper, and some other minor elements. Dental amalgams have a higher thermodynamic tendency for reaction with the oral environment than noble and seminoble cast dental alloys. In addition, these materials receive weaker protection by passivating surface films than implant alloys. However, these materials have shown adequate long-term clinical corrosion resistance. This property has been greatly improved by the transition from low copper amalgams, which contain a corrosion susceptible Sn–Hg structural phase, to high copper amalgams, which contain a more corrosion resistant Sn–Cu phase. Low copper amalgams exhibit breakdown of passivity and suffer from selective corrosion of the tin–mercury phase, which penetrates and weakens the structure. On the other hand, high copper amalgams do not show breakdown in laboratory testing and have demonstrated better clinical performance. The use of dental amalgam in dentistry has been on the decline as a result of concerns regarding the release of small amounts of toxic mercury and due to improvements in the performance of nonmetallic dental composites. Recent reviews on dental alloys and their corrosion behavior can be found in Refs. (24) and (25). Materials for orthodontic applications include cobalt–chromium alloys, titanium alloys, nickel–titanium alloys, which exhibit similar corrosion behavior in dental applications and medical applications.

Ceramic materials were first used in medical devices in the early 1970s. These materials are either crystalline or amorphous, and contain atoms linked by highly directional ionic bonds. Alumina (Al_2O_3) and zirconia (ZrO_2) exhibit high passivation tendencies and resistance to breakdown properties. These materials exhibit better corrosion resistance, hardness, stiffness, wear resistance, and biocompatibility properties than metal alloys. Zirconia and alumina used in medical devices exhibit full-densities and uniformly controlled small grain sizes ($<5\,\mu m$) (26). Full-density ceramics are used in medical devices, because voids may increase stresses and degrade mechanical properties. Ceramics containing uniform small grains are used in order to minimize internal stresses from thermal contraction. In addition, ceramics with small grain sizes exhibit enhanced wear, hardness, and strength properties (27–31). Typical material combinations for ceramic hip prostheses include ceramic-on-ceramic; ceramic-on-metal; and ceramic-on-polymer wear couples.

A ceramic coating material that may provide corrosion resistance to an orthopedic prosthesis is diamond-like carbon (DLC). Diamond-like carbon refers to amorphous carbon materials that contain some component of sp^3-hybridized atoms. Nano- or microcrystalline graphite regions may also be present within the amorphous matrix. Hydrogen-free diamond-like carbon exhibits atomic number densities $>3.19\,g\ atom{\cdot}cm^{-3}$. Hydrogenated diamond-like carbon (HDLC) contains up to 30 atomic percent hydrogen and up to 10 atomic percent oxygen within CH_3 and OCH_3 inclusions, which are surrounded by an amorphous carbon matrix. The density of hydrogenated coatings rarely exceeds $2.2\,g{\cdot}cm^{-3}$. Hydogenated or hydrogen-free diamond-like carbon coatings may provide a medical device with an atomically smooth, low friction, wear resistant, corrosion resistant hermetic seal

between the bulk biomaterial, and the surrounding tissues. Tiainen demonstrated extremely low corrosion rates for diamondlike carbon-coated metals (32). The hydrogen-free diamond-like carbon coated-cobalt–chromium–molybdenum alloy and cobalt–chromium–molybdenum alloy were placed in saline solution equivalent to placement in body fluid for 2 years at a temperature of 37 °C. The DLC-coated cobalt–chromium–molybdenum alloy exhibited 10^5 lower corrosion rate than cobalt–chromium–molybdenum alloy. Similarly, the corrosion rate of DLC-coated titanium–aluminum-vanadium alloy in saline solution has been shown to be extremely low.

Bioactive ceramic materials, which develop a highly adherent interface with bony tissue, have been developed for several medical and dental applications, including coatings for promoting bone ingrowth, grouting agents for hip arthroplasty, and replacements for autologous bone grafts. The most commonly used bioactive ceramics include hydroxyapatite, $Ca_{10}(PO_4)_6(OH)_2$, tricalcium phosphate, $Ca_3(PO_4)_2$, and $Na_2OCaOP_2O_5SiO_2$ glasses (e.g., Bioglass). These materials undergo chemical–biochemical processes, which are dependent on several material properties. For example, 45S5 Bioglass, which contains 45 wt% SiO_2 and 5:1 $CaO:P_2O_5$ ratio, forms SiOH bonds, hydrated silica gel, hydroxyl carbonate apatite layer, matrix, and bone at the material/tissue interface. Materials with high (>60 mol%) SiO_2, low CaO/P_2O_5 ratios, and additions of Al_2O_3, ZrO_2, or TiO_2 are not highly reactive in aqueous media, and do not demonstrate bonding to bone. For example, Bioglass degradation is highly dependent on composition. The dissolution behavior of calcium phosphate ceramics depends on their composition, crystallinity, and processing parameters. For example, materials with larger surface areas (e.g., powders) and smaller grain sizes resorb more rapidly due to preferential degradation at grain boundaries. Phase is another important factor, with alpha-tricalcium phosphate and beta-tricalcium phosphate degrading more slowly than hydroxyapatite. Hydrated forms of calcium phosphate are more soluble than nonhydrated forms. In addition, ionic substitutions affect resorption rate; CO_3^{2-}, Mg^{2+}, and Sr^{2+} increase and F^- decreases biodegradation. Finally, low pH conditions seen in infection and inflammation can result in locally active dissolution processes.

Polymers used in medicine include polyethylene, poly(methyl methacrylate), poly(dimethylsiloxane), poly(tetrafluoroethylene), and poly(ethyleneterephylate). These structures contain primarily covalent atomic bonds, and many undergo several *in vivo* degradation processes. Water, oxygen, and lipids may be absorbed by the polymer, which may result in local swelling. Polyamides avidly absorb lipids and undergo a stress-cracking process known as crazing; these materials may swell up to five volume percent, and can serve as locking inserts for screws. Desorption (leaching) of low molecular weight species can occur due to release of species remaining from fabrication or from chain scission processes, including free radical depolymerization and hydrolysis. Hydrolytic- and enzymatic-based degradation processes are also possible. Wettability also has a prominent effect on the degradation rate of polymers. Degradation of hydrophilic polymers occurs by surface recession, and may resemble uniform corrosion of metals. Hydrophobic polymers may absorb water and other polar species. As a result, the amorphous regions may dissolve preferentially to crystalline ones, increasing the surface area and the effective dissolution rate. A process similar to inter-granular corrosion may result, with abrupt loss of integrity and small particle release.

WEAR

Wear is the loss of material as debris when two materials slide against one another, which may result in abrasion, burnishing, delamination, pitting, scratching, or embedding of debris. The study of wear, friction, and lubrication was integrated in a 1966 British Department of Education and Science report into a new branch of science known as tribology. The term biotribology was coined in 1973 by Dowson to describe wear, friction, and lubrication in biological systems (33). Over the past 30 years, biotribologists have considered the wear properties of orthopedic, dental, cardiovascular, ophthalmic, and urologic devices, including artificial joints, dental restorations, artificial vessels, prosthetic heart valves, and urinary catheters.

Much of biotribology research has focused on orthopedic prostheses, including devices that replace the function of the hip, knee, shoulder, and finger joints. Hip prostheses have provided control of pain and restoration of function for patients with hip disease or trauma, including osteoarthritis, rheumatoid arthritis, osteonecrosis, posttraumatic arthritis, ankylosing spondylitis, bone tumors, and hip fractures. Polymers, metals, ceramics, and composites have been used on the bearing surfaces of orthopedic prostheses. At present, there are three material combinations used in hip prostheses: a metallic head articulating with a polymeric acetabular ceramic cup; a metallic head articulating with a metallic acetabular metallic cup; a ceramic head articulating with a ceramic acetabular polymeric cup.

Osteolysis and aseptic loosening (loosening in the absence of infection) are the major causes of hip prosthesis failure. In 1994, the National Institutes of Health concluded that the major issues limiting hip prosthesis lifetime include the long-term fixation of the acetabular component, biological response due to wear debris, and problems related to revision surgery (34). Although problems with acetabular fixation have been significantly reduced in the intervening years, wear and the biological response to wear debris remain major problems that reduce the longevity of hip prostheses.

Wear may affect the longevity and the function of hip and other orthopedic prostheses. Clinical practices, patient-specific factors, design considerations, materials parameters, and tissue-biomaterial interaction all play significant roles in determining implant wear rates (35). The complex interaction between these parameters makes it difficult to determine a relationship between the in vitro properties of biomaterials and the *in vivo* wear performance for joint prostheses. For example, particles produced by wear may excite both local and systemic inflammatory responses. In addition, the function of prostheses may be affected by the shape changes that are

caused by uneven wear of surfaces. More effective collaboration among clinicians, material scientists, and biologists is necessary to understand the underlying biological, chemical, mechanical, and patient related parameters associated with wear of prostheses.

Wear may occur via adhesive, abrasive, fatigue, or corrosive mechanisms (30–33). The wear process for a given medical device is usually a combination of these mechanisms; however, one mechanism often plays a dominant role. The most important wear mechanism in orthopedic prostheses is adhesive wear. Adhesive wear is caused by adhesive forces that occur at the junction between rough surfaces. Adhesive wear may occur at asperities, or regions of unevenness, on opposing surfaces. Extremely large local stresses and cold welding processes may occur at the junctions between materials. Material may be transferred from one surface to the other as a result of relative motion at the junction. The transferred fragments may be either temporarily or permanently attached to the counterface surface. During this process, the volume of wear material produced is proportional to both the sliding distance acting on the device and the load. The volume of wear materials produced is also inversely proportional to the hardness of the material. For acetabular hip and tibial knee prostheses, adhesive wear is dependent on the large-strain deformation of polyethylene. For acetabular components under multiaxial loading conditions, plastic strain is locally accumulated until a critical strain is reached. Adhesive wear and submicron wear particle release occurs if this critical value is exceeded (30). Although adhesive wear is the most commonly occurring wear mechanism, it is also the most difficult one to prevent.

Abrasive wear takes place when a harder material ploughs into the surface of a softer material, resulting in the removal of material and the formation of depressions on the surface of the softer material. In general, materials that possess higher hardness values exhibit greater resistance to abrasive wear; however, the relationship between resistance to abrasive wear and hardness is not directly proportional. Abrasive wear is called two-body wear when asperities on one surface plough into and cause abrasion on the counterface surface (36). For example, hip prosthesis simulator testing has shown a positive correlation between the surface roughness of the metallic femoral head and the amount of wear damage to the polyethylene acetabular cup. Isolated scratches on a metallic counterface may also participate in abrasive wear. Three-body wear can also occur if hard, loose particles grind between two opposing surfaces that possess similar hardness values. These loose particles may arise from the material surfaces or from the immediate environment, and may become either trapped between the sliding surfaces or embedded within one of the surfaces. For example, metal, polymer, or tissue (e.g., bone) particles embedded in a polyethylene-bearing surface may act to produce third-body wear in orthopedic prostheses. The overall rate of abrasive wear in polyethylene, metal, and ceramic orthopedic prosthesis components depends both on the surface roughness of the materials and the presence of hard third-body particles.

Fatigue wear is caused by the fracture of materials that results from cyclical loading (fatigue) processes. Surface cracks created by fatigue may lead to the generation of wear particles. Cracks deeper within the biomaterial may generate larger particles, in a process known as microcracking. This process typically occurs in metal components; however, has been observed in other materials (e.g., polyethylene) as well. Corrosive wear results from chemical or electrochemical reactions at a wear surface. For example, metals may react with oxygen at a wear surface (oxidation). The resulting oxide may have a lower shear strength than the underlying metal, and may exhibit a more rapid wear rate than the surrounding material. The rate of corrosive wear is governed by the reactivity of the biomaterial, the chemical properties of the implant site, and the mechanical activity of the medical device.

A film or layer of lubricant between the two bearing surfaces can serve to reduce frictional forces and wear. Lubrication can be divided into three regimes: full film (hydrodynamic) lubrication, boundary lubrication, and mixed lubrication. In full film lubrication, the sliding surfaces are entirely separated by a lubricant film that is greater in thickness than the roughness of the surfaces. In boundary lubrication, the surfaces are separated by an incomplete lubricant film, which does not prevent contact by asperities on the surfaces. A mixed lubrication is the one that encompasses aspects of full film and boundary lubrication, in which a region of the two surfaces exhibits boundary lubrication, and the remainder exhibits full film lubrication. The healthy synovial joint provides a low wear and low friction environment, which may exhibit combination of these lubrication modes. Under normal conditions, the hip, knee, and shoulder joints exhibit full film lubrication, in which the two opposing surfaces are entirely separated by a lubricant film of synovial fluid, which carries the load of the joint.

Wear testing is an important consideration during the development of novel biomaterials and medical devices. Any changes in biomaterial or implant design parameters, including composition, processing, and finishing, should be accompanied by studies that confirm that these changes provide either equivalent or improved wear performance to the implant under clinical conditions. As mentioned earlier, asperities on the contact surfaces generally have a significant effect on overall wear performance. In addition, wear has been described as an accumulative process, because overall wear behavior is highly dependent on the material and testing history. An isolated event during a wear test (e.g., the presence of a third-body wear particle) may have a significant impact on the behavior that is observed.

Wear can be assessed in several ways, including which involve changes in shape (dimensions), size, weight of the implant, weight of the debris, or location of radioactive tracers (37). A standard parameter, known as a wear factor, can be used to estimate the wear effects obtained from different wear tests. The wear factor (K) is defined as

$$K = V/LX \tag{4}$$

in which V is the volume of wear (mm^3), L is the applied load (N), and X is the sliding distance (m). Many parameters can influence the results of wear testing,

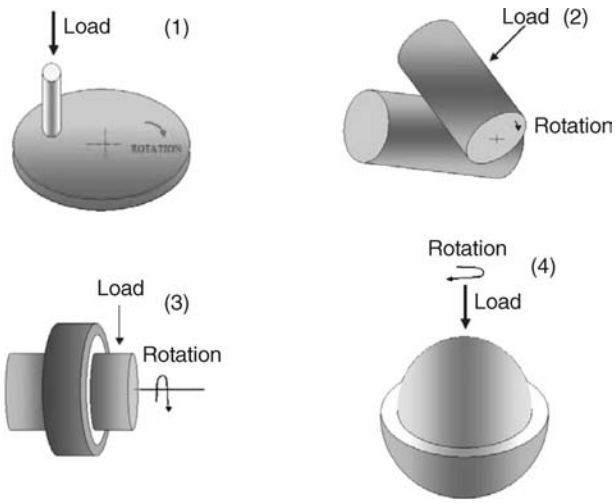


Figure 4. Schematic illustration of geometries in which wear phenomena are likely to occur: (1) pin-on-disk; (2) crossed cylinder; (3) journal bearing; and (4) ball-and-socket bearing (After Ref. 31.)

including test lubricants, test duration, sliding velocity, contact area, alignment, and vibration.

Wear studies fall under three broad categories: (a) screening studies that involve testing of materials with simple geometries under well-controlled conditions; (b) simulator studies that involve testing of partial or complete prostheses; and (c) *in vivo* and retrieval studies of complete implanted devices. Screening studies may provide a basis for comparing novel materials against established materials; however, they can only provide estimations for wear of medical device components. Screening studies involve four general types of geometries: (1) pin-on-disk; (2) crossed cylinder; (3) journal bearing; and (4) spherical or ball and socket bearing. Geometries (3) and (4) are most similar to those encountered in orthopedic protheses (Fig. 4) (35). Simulator studies can be used to assess biomaterials and compare the wear characteristics of materials within a medical device. Various design and material combinations can be examined prior to animal studies and the clinical trials. Clinical assessments and implant retrieval studies also provide useful information for improving biomaterials, medical device design, and manufacturing protocols.

Much of current biotribology research focuses on the relationship between wear performance and biomaterial properties, including composition, processing, and finishing. However, other parameters have significant impact on the wear performance of the prostheses, including surgical and patient factors. The wear performance of polymer, metal, and ceramic biomaterials is discussed below.

Ultrahigh molecular weight polyethylene is commonly used in load-bearing components of total joint prostheses (38–45). The use of a polyethylene/cobalt–chromium wear couple in orthopedic prosthesis was first advocated by Charnley. In many contemporary total hip prosthesis designs, an ultrahigh molecular weight polyethylene acetabular cup slides against a cobalt–chromium alloy femoral ball. Significant numbers of submicron-sized ultrahigh molecular weight polyethylene wear particles are commonly released from these prostheses with each movement of the joint. These particles may remain in the synovial fluid that serves to lubricate the joint (and contribute to third-body wear), embed in prosthesis surfaces, or enter lymphatic circulation. Immune cells (e.g., macrophages) may identify these particles as foreign materials and initiate an inflammatory response, which can lead to rapid bone loss (osteolysis), prosthesis loosening, or bone fracture (39). The volume and size of wear particles are critical factors that affect macrophage activation (40). These biological and physical effects of ultra-high molecular weight polyethylene wear particles are presently the leading cause of long term failure for metal-on-polyethylene hip prostheses (41,42).

Several mechanisms have been proposed to describe wear of ultrahigh molecular weight polyethylene prosthesis components. The wear mechanism of ultrahigh molecular weight polyethylene in hip prostheses has been described by Jasty et al. (43). They found that ultrahigh molecular weight polyethylene surfaces of retrieved implants contained numerous elongated fibrils, which were indicative of large strain deformation. This plastic deformation resulted from strain hardening of the material in the sliding direction and weakening of the material in the transverse direction. Once strain deformation of the surface has occurred, the surface will fragment during the relative motion, and micron- and submicron-sized wear particles will be released. Subsurface cracking, pitting, and delamination caused by oxidative embrittlement and subsurface stresses are responsible for wear of ultra-high molecular weight polyethylene tibial knee inserts.

The wear resistance of ultrahigh molecular weight polyethylene can be improved by reducing the plastic-strain deformation and increasing the oxidization stability (30). The large-strain plastic deformation of ultrahigh molecular weight polyethylene can be diminished by increasing the number of covalent bonds between the long molecular chains of the polymer, which reduces the mobility of the polymer chain and minimizes the creep of the polymer. This process can be achieved by chemical methods (e.g., silane reactions) or, more commonly, by exposing polyethylene to ionizing radiation (46–50). Gamma-ray, e-beam, or X-ray radiation is used to cleave C=C and C=H bonds in polyethylene, which leads to the formation of species with unpaired electrons (free radicals). If the carboncarbon bond is cleaved (chain scission), the polymer molecular weight is reduced. Cross-linking can occur if free radicals from separate chains react with one another, and form an inter-chain covalent bond. If cross-linking occurs as a result of recombination by two radicals cleaved from C=H bonds, it is referred to as an H-type cross-link. If one of the free radicals comes from the cleavage of the C=C bond, it is referred to as a Y-type cross-link. The Y-type cross-linking process can increase the extent of polymer side chain branching (51). The yield of cross-linking processes has been estimated to be three times greater than the yield of chain scission processes for radiation/ultrahigh molecular weight polyethylene interaction. Cross-linking is most significant in amorphous regions of ultrahigh molecular weight polyethylene. An 83% reduction in wear rate has

been reported for ultrahigh molecular weight polyethylene surfaces treated with 5 Mrad radiation (38).

However, not all of the free radicals recombine with other free radicals. In crystalline regions, where the spatial separation between free radicals is large, the residual free radicals become trapped. These species are often confined to the crystalline-amorphous interfaces (52,53). Residual free radicals can cause long-term embrittlement through a series of complex cascade reactions. The residual free radicals first react with oxygen, leading to the formation of oxygen-centered radicals. The oxygen-centered radicals can take a hydrogen atom from a nearby chain to form a hydroperoxide species and another free radical on a chain. This additional free radical can repeat the process by generating another hydroperoxide and forming another free radical on a chain. These unstable species may decay into carbonyl species after exposure to high temperatures or after long periods of time, resulting in lower molecular weights and recrystallization. These processes result in increased stiffness, which is highly undesirable for biotribological applications.

Significant research has been done on reducing the concentration of residual free radicals and limiting the brittleness of irradiated ultrahigh molecular weight polyethylene. One cross-linking postprocessing treatment involved annealing the polymer above its melting transition, which allowed the residual free radicals to be removed through recombination reactions. The polymer recrystallized on cooling; however, the covalent bonds obtained during cross-linking were maintained. Unfortunately, the ultrahigh molecular weight polyethylene exhibited slightly lower crystallinity after this treatment. Another treatment involved annealing the cross-linked polymer at a temperature below the peak melting transition. One advantage of this technique is that a greater degree of crystallinity is retained; however, only a partial reduction in the number of residual free radicals is achieved. Other treatments for residual free radicals include irradiation at room temperature followed by annealing at temperatures below the melting transition; irradiation at room temperature with gamma or electron beams followed by melting; or irradiation at high temperatures followed by melting (34).

The physical properties of the ultrahigh molecular weight polyethylene can be significantly altered by cross-linking and annealing treatments. The effect of these treatments is dependent on the cross-linking parameters (e.g., technique, radiation source, dose, temperature during irradiation) and the annealing parameters (e.g., annealing temperature, annealing time). For example, the ultimate elongation (<45%) and the work to failure for ultrahigh molecular weight polyethylene are reduced as the radiation dose level is increased. Large radiation doses also reduce the yield strength (<30%) and the modulus (<27%) of ultrahigh molecular weight polyethylene (34). In addition, toughness decreases as the radiation dose level is increased, since the energy absorption before failure decreases as the chain mobility is reduced (54).

One alternative to the use of ultrahigh molecular weight polyethylene involves the use of so-called metal-on-metal prostheses, which contain two metallic load-bearing components. The primary motivation for use of these implants is friction; metal-on-metal bearings generate less frictional torque during simulated gait than metal-on-polyethylene bearings (55–59). A stainless steel metal-on-metal hip prosthesis design was attempted by Wiles in 1938. Cobalt–chromium alloy/cobalt–chromium alloy prostheses designs were later developed by McKee and Watson-Ferrar (55,56). Although many of these early cobalt–chromium alloy metal-on-metal hip prostheses failed relatively soon after implantation, others have remained in place for >20 years (35). These first-generation metal-on-metal prostheses were displaced by ultrahigh molecular weight polyethylene/cobalt-chromium alloy prostheses in the 1970s for several reasons, including seizure of the cast metal surfaces (56). In the 1980s, second generation cobalt–chromium alloy/cobalt–chromium alloy prostheses were developed, which have again attracted interest from biomaterials researchers and prosthesis manufacturers (58). Earlier problems with seizing have been minimized through the use of wrought alloys, which are prepared using a thermal-mechanical forming process. Scholes et al. recently demonstrated using a hip simulator system that the mode of lubrication in metal-on-metal hip prostheses is strongly influenced by the diameter of the femoral head and diameter clearance (42). In small diameter joints, the wear rate increased as the diameter of the femoral head was increased. These results were attributed to the development of mixed lubrication in this system.

Alumina and zirconia have also been considered for use in orthopedic prostheses. Alumina exhibits very high hardness and elastic modulus values of 1900 $kgf \cdot mm^{-2}$ (Vickers hardness) and 380 GPa, respectively (60). This material is polished to provide an extremely smooth finish; surface roughness values $<0.005\,\mu m$ are routinely obtained. In addition, alumina surfaces are hydrophilic and may provide prostheses with full film lubrication (35). Fracture toughness and wear resistance can be improved by lowering grain size, increasing grain uniformity, increasing purity, and lowering porosity.

Alumina prostheses have demonstrated wear rates $<1\,mm{\cdot}million^{-1}$ cycles during simulator testing (35). In addition, the *in vivo* wear rate for early alumina-on-alumina hip prostheses was shown to be as low as 1 mm·year (61). However, retrieval studies involving early alumina-on-alumina hip prostheses found high rates of wear on some prostheses. Microseparation of the head and cup was shown to be responsible for this *in vivo* wear behavior (62). Insley et al. examined alumina-on-alumina prostheses with a laboratory simulator, and found that many very small (~40 nm) and some large (100–3000 nm) particles were generated under microseparation conditions (35). Zirconia is harder than alumina, and is used to fabricate smaller components that can withstand higher stresses. Deformation-induced phase transformation has a significant effect on the mechanical properties of zirconia (63). The crystalline phase of a pure zirconium changes from monoclinic to tetragonal during deformation, which is accompanied by volume expansion of ~3–4%. The addition of either yttrium oxide (Y_2O_3) or magnesium oxide (MgO) stabilizes the tetragonal phase at room temperature. However, aging can cause zirconia to return the more stable monoclinic phase, and can limit the lifespan of zirconia

prostheses (64,65). In addition, Sato et al. showed that the tetragonal-monoclinic transformation on the surface of zirconia prostheses can be promoted by the presence of water molecules in the environment (66). The resulting volume change can lead to the generation of surface microcracks and an increase in surface roughness. The failure of some zirconia prostheses has been attributed to this microcracking process. Finally, *in vivo* fracture of some zirconia prostheses has been attributed to variations in sintering.

Thermal oxidation of zirconium alloys has also been used to create biocompatible, corrosion- and wear-resistant surfaces for orthopedic prostheses (67). Wrought zirconium-2.65 weight % niobium alloy contains a two-phase microstructure, which consists of elongated hexagonal alpha-zirconium grains that are bordered by cubic beta-zirconium grains. This material is oxidized for up to 8 h in air at temperatures near 620 °C (the eutectoid temperature). The resulting ~5 μm thick monoclinic ZrO_2 surface contains 40 nm wide × 200 nm long grains that are arranged in a brickwork pattern, which is resilient to grain pull-out and lateral fracture. At the interface between the alloy and the surface oxide, regions of unoxidized niobium in beta-zirconium second-phase grains continue from the alloy into the oxide and serve to anchor the oxide to the alloy. The outermost portion of the oxide surface is burnished to create a smooth bearing surface. The oxide surface provides excellent wear behavior against polyethylene components, with reduced wear particle generation and inflammation.

Diamond-like carbon coatings on orthopedic prostheses can exhibit a wide range of elastic modulus and hardness values, which can be correlated with the fraction of sp^3-hybridized atoms within the coating (68–71). Collins et al. developed a relation between sp^3 fraction and Vickers hardness values for hydrogen-free diamond-like carbon coatings. They found that an sp^3 fraction of 10% corresponded to a hardness value of 2000–3000 Hv, an sp^3 fraction of 50% corresponded to a hardness value of 7000-8000 Hv, and an sp^3 fraction of 100% corresponded to a hardness value of 10,000 Hv (72). Schneider et al. found that hydrogen-free and diamond-like carbon films with sp^3 fractions between 0 and 90% provided elastic modulus values between 300 and 800 GPa. In contrast, typical hardness and elastic modulus values for hydrogenated diamond-like carbon films are <17 and <200 GPa, respectively (73).

The coefficients of friction values for diamond-like carbon coatings depend on ambient humidity, topology, and sliding partner (74). The most important parameter determining the coefficient of friction for hydrogenated diamond-like carbon thin films is relative humidity. Friction values for hydrogenated diamond-like carbon films can be as low as 0.01–0.3 in vacuum conditions, but greatly increase under humid conditions due to incomplete formation of the graphitic transfer surface. This variation in coefficient of friction values can be correlated with hydrogen/carbon ratio in the precursor material. As the hydrogen content in the precursor material increases, the friction coefficient demonstrates a greater positive correlation with ambient humidity. For example, hydrogenated diamond-like carbon films produced from hydrogen-diluted methane demonstrate lower friction coefficients under high humidity conditions than other hydrogenated diamond-like carbon films. On the other hand, hydrogen-free diamond-like carbon thin films maintain low friction coefficients (<0.1) under low and high humidity conditions (75,76).

The combination of high hardness and low coefficient of friction values allows diamond-like carbon coatings to provide significant wear protection to a bulk implant material (71). Hirvonen et al. found that the wear resistance of diamond-like carbon coatings is superior to that of silicon carbide, tungsten carbide–cobalt, silicon nitride, or alumina by factors of 40, 60, 230, and 290, respectively (77). Hydrogen-free diamond-like carbon thin films exhibit wear rates of 10^{-9} $mm^3 \cdot N^{-1} \cdot m^{-1}$, these values are ~100 times lower than those for hydrogenated diamond-like carbon thin films (10^{-7} $mm^3 \cdot N^{-1} \cdot m^{-1}$) under similar testing conditions (78). Many diamond-like carbon substrate materials are significantly softer than the diamond-like carbon coatings; high contact pressures can initiate substrate deformation and coating failure. Nitriding processes can harden the substrate surface, reduce subsurface deformation, and extend diamond-like carbon coating lifetimes (79).

The friction and wear properties of diamond-like carbon-coated metal hip prostheses against diamond-like carbon-coated polyethylene cups have been determined using both screening and simulator techniques (80). For example, Tiainen et al. demonstrated extremely low coefficients of friction for prostheses coated with hydrogen-free diamond-like carbon using a pulsed arc discharge method. They demonstrated coefficients of friction for diamond-like carbon/diamond-like carbon and metal–metal pairs of 0.05 and 0.14, respectively. In addition, they found that wear rate in the diamond-like carbon-coated metal/diamond-like carbon-coated metal wear couple was 10^5–10^6 times lower than that observed in conventional metal–polyethylene and metal–metal wear couples. They also observed that wear of polyethylene in the diamond-like carbon-coated metal/ultrahigh molecular weight polyethylene wear couple was 10–600 times lower than that observed in conventional metal/ultrahigh molecular weight polyethylene wear couples. On the other hand, other investigators have found little difference in wear rates between diamond-like carbon-coated prosthesis materials and conventional prosthesis materials. For example, Sheeja et al. found little difference in wear rates between cobalt-chromium–molybdenum alloy/ultrahigh molecular weight polyethylene and multilayered diamond-like carbon-coated cobalt–chromium–molybdenum alloy/ultrahigh molecular weight polyethylene wear couples (81,82). The seemingly contradictory results suggest other factors, such as use of lubricant, may play a significant role in determining wear rates (83–85). For example, physiologic lubricants may not allow graphitic layers to form on the surfaces of the test materials. In addition, diamond-like carbon coatings that contain particulates and pits may increase adhesive wear (86).

Adhesion of diamond-like carbon coatings is dependent on several factors, including film stress, film–substrate chemical bonding, and substrate topology (87,88). Large internal compressive stresses (as high as 10 GPa) are typically observed in

diamond-like carbon coatings. These stresses limit maximum diamond-like carbon coating thickness to 0.1–0.2 mm and prevent widespread medical use. Lifshitz et al. attributed these stresses to subplantation (low energy subsurface implantation) of carbon ions during coating formation (89). They suggested that carbon ions with energies between 10 and 1000 eV undergo shallow implantation to depths of 1–10 nm during growth of diamond-like carbon coatings. Carbon species are trapped in subsurface sites due to restricted mobility, leading to the development of very large internal compressive stresses.

Diamond-like carbon can be alloyed with metals in order to reduce internal compressive stresses and promote specific biological responses (90,91). Diamond-like carbon–metal composite coatings retain hardness and wear properties similar to those of unalloyed diamond-like carbon films, and exhibit excellent adhesion to metal alloy substrates (Fig. 5). The metal component can provide additional biological functionality to the implant surface; for example, silver has been shown to possess a wide antimicrobial spectrum against a broad range of Gram-negative bacteria (including *Pseudomonas aeruginosa*),

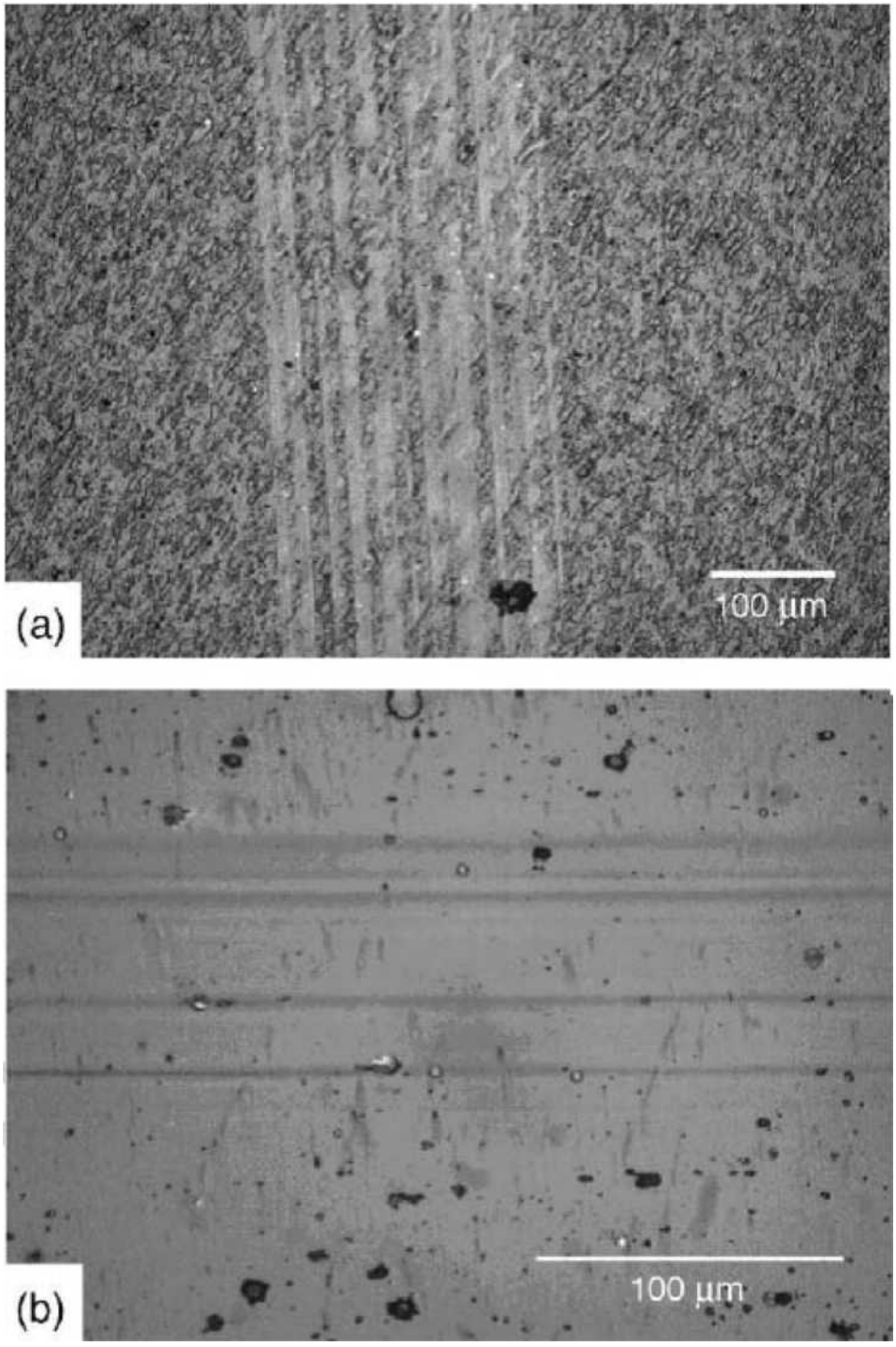


Figure 5. Wear testing of diamond-like carbon–metal composite coatings using a linear tribometer (screening wear test). (a) Wear track of functionally gradient diamond-like carbon–silver composite film after 10,000 cycles, (b) Wear track of functionally gradient diamond-like carbon–titanium composite film after 10,000 cycles.

Gram-positive bacteria (including methicillin-resistant *Staphylococcus aureus*), fungi, viruses, and yeasts. Films containing both silver and platinum may demonstrate enhanced antimicrobial activity due to formation of a galvanic couple that accelerates silver ion release. Narayan et al. showed that diamond-like carbon–silver–platinum nanocomposite films reduce bacterial colonization rates by 90% compared to uncoated silicon substrates (92).

Corrosion and wear are critical parameters that affect overall success of a biomaterial or a medical device design. The complicated interaction between material-, device-, surgical-, patient-specific parameters has made it difficult to predict clinical behavior of implanted medical devices. In addition, the large variation in measurement techniques for corrosion and wear has led problems in interpreting and comparing the work performed in the biomaterials community. If these issues can be successfully resolved, significant advances in these areas may be achieved in the coming years.

BIBLIOGRAPHY

Cited References

1. Burke DR. The Composition and Function of Body Fluids, 3rd ed. St. Louis: C. V. Mosby; 1980.
2. Lentner C, editor. Geigy Scientific Tables, Vol. 1: Units of Measurement, Body Fluids, Composition of the Body, Nutrition. Basel: CIBA-Geigy; 1981.
3. Lentner C, editor. Geigy Scientific Tables, Vol. 3: Physical Chemistry, Composition of Blood, Hematology, Somatometric Data. Basel: CIBA-Geigy; 1984.
4. ASTM, Annual Book of ASTM Standards. West Conshohocken (PA): ASTM. International; 1977.
5. Pilliar RM. Metals and irthopaedic implants — past successes, present limitations, future challenges. Shrivastava S, editor. Proceedings of the Materials & Processes for Medical Devices Conference. Materials Park (OH): ASM International; 2004.
6. Hoar TP, Mears DC. Corrosion-resistant alloys in chloride solutions: materials for surgical implants. Proc R Soc London 1966;294:486–510.
7. Sury P, Semlitsch M. Corrosion behavior of cast and forged cobalt-based alloys for double-alloy joint endoprostheses. J Biomed Mater Res 1978;12:723–741.
8. Steinemann S. Corrosion of surgical implants – *in vivo* and *in vitro* tests. In: Winder GD, editor. Evaluation of Biomaterials. Chichester: John Wiley & Sons Inc.; 1980.
9. Cahoon JR, Bandyopadhya R, Tennese L. The concept of protection potential applied to the corrosion of metallic orthopedic implants. J Biomed Mater Res 1975;9:259–264.
10. Bandy CR. Effects of composition on the electrochemical behavior of austenitic stainless steel in Ringer's solution. Corrosion (Houston) 1977;33:204–208.
11. Scales JT, Winter GD, Shirley HT. Corrosion of orthopaedic implants. J Bone Joint Surg 1959;41B:810–820.
12. Merritt K, Brown SA. Release of hexavalent chromium from corrosion of stainless steel and cobalt-chromium alloys. J Biomed Mater Res 1995;29:627–633.
13. McKellop HA, Sarmiento A, Brien W, Park SH. Interface corrosion of a modular head total hip prosthesis. J Arthroplasty 1992;7:291–294.
14. Brown SA, et al. Fretting corrosion accelerates crevice corrosion of modular hip tapers. J Appl Biomater 1995;6:19–26.
15. Kawale JS, Brown SA, Payer JH, Merritt K. Mixed-metal fretting corrosion of Ti6Al4V and wrought cobalt alloy. J Biomed Mater Res 1995;29:867–873.
16. Metikos-Hukovic M, Kwokal A, Piljac J. The influence of niobium and vanadium on passivity of titanium-based implants in physiological solution. Biomaterials 2003;24: 3765–3775.
17. Cutright DE, et al. Tissue reaction to nitinol wire alloy. Oral Surg Oral Med Oral Pathol 1973;35:578–584.
18. Kapanen A, Ryhänen J, Danilov A, Tuukkanen J. Effect of nickel-titanium shape memory metal alloy on bone formation. Biomaterials 2001;22:2475–2480.
19. Ryhänen J, et al. Bone healing and mineralization, implant corrosion, and trace metals after nickel-titanium shape memory metal intramedullary fixation. J Biomed Mater Res 1999;47:472–480.
20. Villermaux F, et al. Corrosion resistance improvement of NiTi osteosynthesis staples by plasma polymerized tetrafluorethylene coating. Biomed Mater Eng 1996;6:241–254.
21. Rondelli G, Vincentini B. Localized corrosion behavior in human body fluids of commercial NiTi orthodontic wires. Biomaterials 1999;20:785–792.
22. Carroll W, Kelly M, Brien B. Corrosion behavior of Nitinol wires in body fluid environment. Int Conf Shape Memory Superelastic Technol 1999; 240–249.
23. Montero-Ocampo C, Lopez H, Salinas RA. Effect of compressive straining on corrosion resistance of a shape memory Ni-Ti alloy in Ringer's solution. J Biomed Mater Res 1996;32: 583–591.
24. Shabalovskaya SA. Surface, corrosion and biocompatibility aspects of Nitinol as an implant material. Bio-Med Mater Eng 2002;12:69–109.
25. Niinomi M. Recent research and development in titanium alloys for biomedical applications and healthcare goods. Sci Technol Adv Mat 2003;4:445–454.
26. Senda T, Yasuda E, Kaji M, Bradt RC. Effect of Grain Size on the Sliding Wear and Friction of Alumina at Elevated Temperatures. J Am Ceram Soc 1999;82:1505–1511.
27. Dogan CP, Hawk JA. Role of composition and microstructure in the abrasive wear of high-alumina ceramics. Wear 1999; 225:1050–1058.
28. Rodríguez J, et al. Sliding wear of alumina/silicon carbide nanocomposites. J Am Ceram Soc 1999;55:2252–2306.
29. Webster TJ, Siegel RW, Bizios R. Osteoblast adhesion on nanophase ceramics. Biomaterials 1999;20:1221–1227.
30. Morsi K, Keshavan H, Bal S. Hot pressing of graded ultrafine-grained alumina bioceramics. Mater Sci Eng A 2004;386:384–389.
31. Kingery WD. Introduction to Ceramics. New York: John Wiley & Sons Inc.; 1976.
32. Tiainen VM. Amorphous Carbon as a Bio-mechanical coating-mechanical properties and biological applications. Diamond Related Mater 2001;10:153–160.
33. Hutchings IM. Biotribology-A Personal View. In: Hutchings IM, editor. Friction, Lubrication and Wear of Artificial Joints. Bury St. Edmunds (UK): Professional Engineering Publishing Ltd.; 2003.
34. Wright TM, Goodman SB, editors. Implant Wear in Total Joint Replacement: Clinical and Biologic Issues, Material and Design Considerations. Rosemont (IL): American Academy of Orthopaedic Surgeons; 2001.
35. Hutchings IM, editor. Friction, Lubrication and Wear of Artificial Joints. Bury St. Edmunds (UK): Professional Engineering Publishing Ltd; 2003.
36. Schmalzreid TP, Callaghn JJ. Current concepts review: Wear in total hip and knee replacement. J Bone Joint Surg Am 1999;81:115–136.

37. Clarke IC, McKellop HA. Wear Testing. In: von Recum AF, editor. Handbook of biomaterials evaluation: scientific, technical, and clinical testing of implant materials. New York: Macmillan; 1986.
38. McKellop H, Shen FW, DiMaio W, Lancaster JG. Wear of gamma-crosslinked polyethylene acetabular cups against roughened femoral balls. Clin Orthop 1999;369:73–82.
39. Unsworth A, Dowson D, Wright V. Some new evidence on human joint lubrication. Ann Rheum Dis 1975;34:277–285.
40. Green TR, et al. Effect of size and dose on bone resorption activity of macrophages by *in vitro* clinically relevant ultra high molecular weight polyethylene particles. J Biomed Mater Res 2000;B53:490–497.
41. Ingham E, Fisher J. Biological reactions to wear debris in total joint replacement. Proc Inst Mech Eng H 2000;214:21–37.
42. Scholes C, Unsworth A. Comparison of friction and lubrication of different hip prostheses. Proc Inst Mech Eng H 2000;214:49–57.
43. Jasty MJ, et al. Wear of polyethyene acetabular components in total hip arthroplasty: an analysis of 128 components retrieved at autopsy or revision operation. J Bone Joint Surg Am 1977;79:349–358.
44. Charnley JC. Arthroplasty of the hip: a new operation. Lancet 1961;280:1129–1132.
45. Charnley JC. Tissue reaction to polytetrafluorethylene. Lancet 1963;285:1379.
46. Collier JP, et al. Overview of polyethylene as a bearing material: comparison of sterilization methods. Clin Orthop Rel Res 1996;333:76–86.
47. Wang A, Sun DC, Stark C, Dumbleton JH. Wear Mechanisms of UHMWPE in total joint replacements. Wear 1998;181–183:241–249.
48. Collier JP, et al. Results of implant retrieval from postmortem specimens in patients with well-function, long term total hip replacement. Clin Orthop 1992;274:97–112.
49. Cameron HU. Tibial component wear in total knee replacement. Clin Orthop 1994;309:29–32.
50. Atkinson JR, Cicek RZ. Silane crosslinked polyethylene for prosthetic applications: Part II. Creep and wear behavior and a preliminary molding test. Biomaterials 1984; 5:326–335.
51. Muratoglu OK, et al. Larger diameter femoral heads used in conjunciton with a highly cross-linked ultra high molecular weight polyethylene: A New Concept. J Arthroplasty 2001;16: 24–30.
52. Pearson RW. Mechanism of the radiation crosslinking of polyethylene. J Polym Sci 1957;25:189–200.
53. Zhu QR, Horii F, Kitamaru R, Yamaoka H. C-13-NMR study of cross-linking and long-chain branching in linear polyethylene induced by Co-60 gamma-ray irradiation at different temperatures. J Polym Sci, Part A: Polym Chem 1990;28: 2741–2751.
54. Premnath V, Harris WH, Jasty M, Merrill EW. Gamma sterilization of UHMWPE articular implants: an analysis of the oxidation problem. Biomaterials 1996;17:1741–1753.
55. Steinberg DR, Steinberg ME. The early history of arthroplasty in the United States. Clin Orthop Relat Res 2000;374:55–89.
56. McKee GK, Watson-Farrar, J. Replacement of arthritic hip by the McKee-Farrar prostheses. J Bone Joint Surg 1966;48B: 245–259.
57. Wimmer MA, et al. The acting wear mechanisms on metal-on-metal hip joint bearings: in vitro results. Wear 2001; 250:129–139.
58. Dorr LD, et al. Total hip arthroplasty with use of the metasul metal-on-metal articulation: 4–7-year results. J Bone Joint Surg Am 2000;82:789–798.
59. Poggie RA, Afflitto R, St John K. The wear performance of precision Co–Cr–Mo alloy metal-on-metal hip bearings. Proc Conf Trans 12th Ann Int Symp Technol Arthroplasty 1999;11: 1–2.
60. Boutin P, et al. The use of dense alumina- alumina ceramic combination in total hip replacement. M J Biomed Mater Res 1988;22:1203–1232.
61. Willmann G. Development in medical-grade alumina during the past two decades. J Mater Process Technol 1996;56:168–176.
62. Nevelos J, et al. Microseparation of the centers of alumina-alumna artificial hip joints during simulator testing produces clinically relevant wear and patterns. J Arthroplasty 2000;15: 793–795.
63. Christel P, et al. Mechanical properties and short-term In vivo evaluation of yttrium-oxide-partially-stabilized zirconia. J Biomed Mater Res 1989;23:45–61.
64. Chevalier J, et al. Critical effect of cubic phase on aging in 3 mol% yttria-stabilized zirconia ceramics for hip replacement prosthesis. Biomaterials 2004;25:5539–5545.
65. Gremillard L, et al. Modeling the aging kinetics of zirconia ceramics. J Eur Ceram Soc 2004;24:3483–3489.
66. Sato T, Ohtaki S, Endo T, Shimada M. Science and technology of zirconia. Advances in Ceramics. Westerville (OH): American Ceramic Society; 1988.
67. Hobbs LW, et al. Oxidation microstructures and interfaces in the oxidized zirconium knee. Int J Appl Ceram Technol 2005;2:221–246.
68. Enke K, Dimigen H, Hubsch H. Frictional properties of diamondlike carbon layers. Appl Phys Lett 1980;36:291–292.
69. Pharr GM, et al. Hardness, elastic modulus, and structure of very hard carbon films produced by cathodic-arc deposition with substrate pulse-biasing. Appl Phys Lett 1996;68:779–781.
70. Ronkainen H, Varjus S, Koskinen J, Holmberg K. Differentiating the tribological performance of hydrogenated and hydrogen-free DLC coatings. Wear 2001;249:260–266.
71. Holmberg K, Mathews A. Coatings tribology: a concept, critical aspects, and future directions. Thin Solid Films 1994;253:173–178.
72. Collins CB, et al. Noncrystalline films with the chemistry, bonding, and properties of diamond. J Vac Sci Technol B 1993;11:1936–1941.
73. Schneider D, Schwarz T, Scheibe HJ, Panzner M. Nondestructive evaluation of diamond and diamond-like carbon films by laser induced surface acoustic waves. Thin Solid Films 1997;295:107–116.
74. Erdemir A, Bindal C, Pagan J, Wilbur P. Characterization of Transfer layers on Surfaces Sliding Against Diamond-like Hydrocarbon Films in Dry Nitrogen. Surf Coatings Technol 1995;76–77:559–563.
75. Liu Y, Erdemir A, Meletis EI. An investigation of the relationship between graphitization and frictional behavior of DLC coatings. Surface Coatings Technol 1996;86:564–568.
76. Oguri K, Arai T. Friction mechanisms of Diamond-like carbon with silicon coatings formed by plasma-assisted chemical vapor-deposition. J Mater Res 1992;7:1313–1316.
77. Hirvonen JP, Koskinen J, Lappalainen R, Anttila A. Preparation and properties of high density hydrogen free hard carbon films with direct ion beam or arc discharge deposition. Mater Sci Forum 1990;52:197.
78. Voevodin AA, Donley MS, Zabinski JS, Bultman JE. Mechanical and tribological properties of diamond-like carbon coatings prepared by pulsed laser deposition. Surface Coatings Technol 1995;77:534–539.
79. Voevodin AA, Phelps AW, Zabinski JS, Donley MS. Friction induced phase transformation of pulsed laser deposited diamondlike carbon. Diamond Related Mater 1996;5:1264–1269.

80. Santavirta S, Lappalainen R, Heinonen H, Anttila A. Some relevant issues related to the use of amorphous diamond coatings for medical applications. Diamond Related Mater 1998;7:482–485.
81. Sheeja D, et al. Mechanical and tribological characterization of diamond-like carbon coatings on orthopedic materials. Diamond Related Mater 2001;10:1043–1048.
82. Sheeja D, Tay BK, Lau SP, Nung LN. Tribological characterization of diamond-like carbon coatings on Co-Cr-Mo alloy for orthopaedic applications. Surface Coatings Technol 2001;146: 410–416.
83. Ahlroos T, Saikko V. Wear of prosthetic joint materials in various lubricants. Wear 1997;211:113–119.
84. Saikko V, Ahlroos T. Phospholipids as boundary lubricants in wear tests of prosthetic joint materials. Wear 1997;207:86–91.
85. Affatato S, Frigo M, Toni A. An *in vitro* investigation of diamond-like carbon as a femoral head coating. J Biomed Mater Res 2000;53:221–226.
86. Dong H, Shi W, Bell T. Potential of improving tribological performance of UHMWPE by engineering the Ti6Al4V counterfaces. Wear 1999;229:146–153.
87. Morshed MM, McNamara BP, Cameron DC, Hashmi MSJ. Effect of surface treatment on the adhesion of DLC film on 316L stainless steel. Surface Coatings Technol 2003;163:541–545.
88. Schwan J, et al. Stress-induced formation of high-density amorphous carbon thin films. J Appl Phys 1997;82:6024–6030.
89. Lifshitz Y, et al. Growth mechanisms of DLC films from C+ ions- experimental studies. Diamond Related Mater 1995;4: 318–323.
90. Narayan RJ, Scholvin D. Nanostructured carbon-metal composite films. J Vac Sci Technol B 2005;23:1041–1046.
91. Narayan RJ. Pulsed laser deposition of functionally gradient diamondlike carbon-metal nanocomposites. Diamond Related Mater 2005;14(8):1319–1330.
92. Narayan RJ, et al. Antimicrobial properties of diamond-like carbon-silver-platinum nanocomposite thin films. J Mater Eng Perform 2005;14:435–440.

See also BIOMATERIALS, TESTING AND STRUCTURAL PROPERTIES OF; HEART VALVE PROSTHESES; HIP JOINTS, ARTIFICIAL.

BIOMATERIALS FOR DENTISTRY

STEVE ARMSTRONG
University of Iowa

INTRODUCTION

Gold has been used for dental purposes for at least 2500 years; the fabrication of gold crowns and bridgework flourished in Etruria and Rome as early as 700–500 BC. Gold leaf came into use during the fifteenth century for the restoration of carious teeth. Restorative materials and techniques continued to develop through the nineteenth century including the use of waxes, fused porcelain, "silver paste" amalgam, cements, vulcanite, the angle handpiece, and a gold inlay casting machine. A rapid development in materials and instrumentation has occurred since the 1950s, to include the high speed handpiece, steel and diamond cutting instruments, adhesive techniques to metal, ceramics, enamel and dentin, resin-based composites, glass ionomers, base-metal alloys for partial dentures, metal-ceramic systems, high-strength all ceramic structures, and titanium alloys for dental implants. This increasing complexity and body of knowledge has lead to the establishment of uniform material standards and the recognition of the science of dental materials as a distinct and essential branch of dentistry.

Biomaterials are used in the oral cavity either to restore function, comfort, or aesthetics caused by developmental disorders, disease, or trauma. More elective procedures are being requested and performed purely for aesthetic purposes as the incidence of caries has dropped in certain population groups and as patients have become more aware of various restorative or cosmetic options. However, the replacement of diseased tooth structure or missing teeth accounts for the bulk of work in restorative dentistry. The instruments and materials used in the surgical aspects of oral, maxillofacial and periodontal surgery have much in common with medicine. This article will focus on those commonly used materials in the restoration of individual teeth or the replacement of missing teeth.

Restorative materials include noble and base metal alloys, resin-based composites (RBCs), glass ionomers, ceramics, acrylics, and amalgam alloys. Techniques to apply these materials include both direct and indirect approaches. Materials or "fillings" can be directly placed in a prepared cavity by the use of adhesives and/or retentive-type preparations. Full or partial coverage crowns, bridges, and dentures are fabricated indirectly by dental laboratories or computer aided milling machines and then attached or cemented into the mouth for the coverage of missing or weakened tooth structure or the replacement of missing teeth. Various forms of ceramic or metallic implants can be placed into the upper or lower jaw bones to serve as tooth root substitutes upon which a prosthesis is attached to replace missing teeth. Auxiliary materials, such as waxes, gypsum, and impression materials are also utilized during clinical and laboratory steps but will not be covered in this article.

DIRECTLY PLACED RESTORATIVE MATERIALS: 'FILLINGS'

Amalgam

Dental amalgam has been very successfully used in dentistry for 150 years and is one of the most technique insensitive dental restorative materials available. Dental amalgams are inexpensive and have demonstrated a relatively long service life. The disadvantages of amalgam are the silver color and the presence of mercury. The presence of mercury requires regulatory control of wastewater effluent and has raised unsubstantiated health concerns regarding mercury toxicity to the individual patient.

Dental amalgam is a mixture of mercury and a solid metal alloy of silver, tin, copper, and sometimes zinc, palladium, indium, and selenium. Once the mercury and alloy is mixed, the plastic mass is condensed into the prepared cavity and carved to required form before hardening. The alloy particles are microspheres of various sizes, irregular lathe-cut particles or mixtures of the two. "High

copper" alloys (13–30%) have essentially replaced the low copper alloys of the past. These high copper alloys, along with the addition of zinc for manufacturing procedures, have improved early clinical strength, lowered creep, and improved corrosion resistance. The mixing of mercury and the alloy is referred to as trituration or amalgamation. A surface reaction occurs between the alloy and liquid mercury that binds the unreacted particles together by a surrounding matrix of reaction products. Increasing the copper content eliminates most of the weakest and most corrosive phase ($Sn_{7-8}Hg$) from the setting reaction.

After the carious lesion is removed, the plastic dental amalgam is condensed into the prepared cavity before hardening and subsequently retained by the mechanical resistance and retention form of the surgically prepared cavity. An adhesive liner is not required. The material slowly corrodes and the corrosion products "self-seal" the margin between tooth and amalgam, thereby protecting the tooth from leakage of oral fluids and bacteria and their byproducts. Dental amalgam is brittle and undergoes creep at mouth temperature, which can lead to marginal or bulk fracture and clinical failure. However, if the cavity is well designed and the amalgam placed with technical competence, many years of service should be expected. A vast number of studies have shown the safety and efficacy of dental amalgam. When a dentist is faced with a patient's request to remove amalgam fillings due to a claimed medical malady, the dentist is professionally obligated to explain that the possibility of their medical condition(s) being related to the presence of dental amalgam fillings is extremely remote. These patients typically face a complex problem with biological, psychological, and social components unrelated to mercury intake.

Resin Composites

Modern day resin-based composites placed with dental adhesives have replaced the silicates and acrylic resins of the past and are now widely used throughout dentistry. In addition to the treatment of decay and trauma, RBCs are used in aesthetic or cosmetic dentistry procedures due to their versatility and conservative nature. Discolored, misshapen, or misaligned dentition can be aesthetically treated with cosmetic "bonding". The RBCs are composed of four main components: (*1*) a continuous organic polymer matrix, (*2*) a dispersion of inorganic filler particles, (*3*) silane coupling agents to bind the filler particles with the polymer matrix, and (*4*) an initiator–accelerator system. They also contain various pigments for matching tooth shades and ultraviolet (UV) absorbers to minimize oxidative color changes. The two most common oligomers are the dimethacrylates 2,2- bis[4(2-hydroxy-3-metharcyloyloxy-propyloxy)-phenyl] (bis-GMA) and urethane dimethacrylate (UDMA). Diluents such as triethylene glycol dimethacrylate (TEGDMA) are added to reduce the viscosity for the addition of filler and to obtain clinically acceptable handling properties. The inorganic filler particles can be of borosilicate, lithium aluminum silicate, barium aluminum silicate, strontium or zinc glasses, quartz, or colloidal silica. The combination of relatively larger glass or quartz particles and a significant addition of colloidal silica are referred to as a "hybrid". A useful classification method is by filler particle size (Table 1) with minifill hybrids and microfills being the most popular types. Recently, using a proprietary process, manufacturers have been able to produce a smaller average silica particle size (0.02 μm) as compared with traditional microfills (0.04 μm). Marketed as "nanofills", these smaller silica particles are produced in a nondrying method thereby avoiding agglomeration due to physical forces and thusly enabling a higher degree of filler loading. The microfill RBCs polish to the most enamel-like surface, but lack the strength of the hybrids due to the lower filler volume. Newer formulations of minifill hybrids can be used as "universal" RBCs possessing both the strength for posterior chewing forces and acceptable surface finish for use in aesthetic anterior regions. Clinical studies showed that the long-term wear resistance of RBC restorations placed on posterior teeth is still inferior to the dental amalgam restorations.

Table 1. Classification of Resin-Based Composites by Filler Particle Size

Class	Particle Size, μm	Filler Loading, vol%
Macrofill	>10	50–70
Midifill	1–10	50–70
Minifill	0.1–1	50–70
Microfill	0.01–0.1	20–50

Polymerization occurs through either a self-cured free radical initiation when a peroxide–amine system is mixed or through light-activated free radical initiation when a diketone–amine system is exposed to blue light. The photoactivator, most commonly camphoroquinone, is added in small amounts, which forms a free radical when exposed to 467-nm blue light. Dual-cure varieties of RBCs are available as well. Halogen light-curing units are most commonly used, but several other light curing units are currently being marketed to include: plasma-arc, laser and light-emitting diodes. To insure proper polymerization care must be taken to match the unit's spectral emission and the RBCs spectral requirements. One aspect that all RBCs currently share is 2–4% volumetric shrinkage of the continuous polymeric network upon polymerization. Shrinkage induces residual stress that can disrupt the adhesive bond between the RBC and the tooth structure, damage enamel or the RBC. Incremental placement of light-cured RBCs can help to reduce the net effect of this shrinkage stress. Manufacturer's are currently working to develop no- or low shrinkage RBCs; several approaches are noted: (*1*) addition of ring-opening monomers that expand upon polymerization (spiroorthocarbonates, oxiranes, vinylcyclopropanes), (*2*) low shrinkage cyclopolymerizable di- and multifunctional oligomers synthesized through the reaction of acrylates and formaldehyde, and (*3*) the addition of strain-absorbing polybutadiene rubber polymer adsorbed onto the fumed silica.

Variations in filler loading, viscosity, and polymerization initiating systems allow RBCs to be used in a wide variety of clinical situations, to include: sealants, cements, crown core buildups, so-called flowables and packables,

provisional restorations, and in a variety of laboratory processed RBCs for adhesive cementation as inlays, onlays, crowns and veneers. Several manufacturer's have promoted fiber reinforced RBCs for use as bridges, as well, but these lack convincing clinical data for their recommended usage.

Glass Ionomers

Smith (1968) developed glass ionomer cements (GIC) by combining the polyacrylic acid from polycarboxylate cement, which is strongly adhesive to tooth structure, and the aluminosilicate glass from the fluoride-containing silicate cements. The GICs form a true chemical bond to the tooth mineral (hydroxyapatite) by ionic bonding between calcium and carboxylic ions and act as chemotherapeutic agents in the treatment and prevention of dental caries through the release of fluoride.

The GIC are composed of a basic ion-leachable aluminosilicate glass powders that, upon exposure to water-soluble homopolymer or copolymers of alkenoic acids, form a matrix of continuous polysalts (polyalkenoates) surrounded by partially solubilized glass filler particles. The clinical placement technique must account for the relatively slow-setting reaction and moisture sensitivity. Fluoride easily passes in and out of the matrix without any degradative effects due to the substitution of carboxyl ions for fluoride within the salt matrix. Fluoride "recharging" has been clearly demonstrated in vitro to prevent the demineralization of tooth structure at the margins adjacent to the GIC under an artificial caries challenge. In vivo evidence is not as clearly demonstrated, but when evaluating clinical data from high caries risk cohort populations the anticariogenic effect of GIC is elucidated.

Since their development GIC has been modified in a number of different ways to improve their clinical handling properties and durability. A significant development was the addition of water soluble monomers, for example, HEMA, and the grafting of methacrylate side groups on the polyacid polymer. By the addition of visible light initiator–accelerator systems these resin-modified glass ionomers (RMGI) can be command set with a light curing unit while also self-curing through the acid–base reaction. These improvements to the conventional GIC and RMGI have made these materials widely used as restorative materials. The self-adhesive and self-cure properties of GIC, along with improvements in strength have allowed these materials to be used in nontraditional field situations without the luxuries of electricity or modern equipment. Auxiliary personnel have been trained to remove caries with sharp hand instruments with the cavity then filled with a heavier filled GIC. This treatment has aided thousands through a technique termed Atraumatic Restorative Treatment or "ART". These materials are also used as cements and liners.

Adhesives, Cements, and Liners

Adhesive dentistry has become increasingly important as the use of dental amalgam and direct compacted gold foil and cemented gold restorations has declined. Unlike dental amalgam, RBCs require an adhesive liner for placement and retention. A durable bond of RBCs to enamel can be accomplished by first cleaning and demineralizing the surface with a 30–40% phosphoric acid, followed by a polymerizable methacrylate monomer [bis(GMA), UDMA, TEGDMA], which diffuses into the porosities created by the acid etching. However, bonding to dentin is a much greater challenge due to the compositional differences in dentin relative to enamel and its extremely variable clinical presentation. Dentin contains less inorganic components and more organic components and water than enamel. Dentin is made permeable by tubules that travel from the dental pulp through the dentin to the coronal enamel.

Similar to enamel, the dentin is treated with an acid that removes any smear layer and exposes the collagen fibers by demineralizing the surface. An adhesive primer containing a hydrophilic solvent (water, acetone, ethanol, or HEMA) and an amphipathic monomer (hydrophilic–hydrophobic functionalities) then penetrates the exposed collagen network. After the solvent is evaporated from the primed surface, an adhesive monomer is applied that attaches to the hydrophobic functionality of the primer to create a wetted surface for subsequent copolymerization with the RBC. This bonding process is also approached with so-called "self-etch" adhesives. The initial acid application is eliminated and a water-soluble acidic polymer is included in the primer to simultaneously demineralize the tooth surface while penetrating with the adhesive monomers and oligomers. Regardless of the approach, the adhesive liner penetrates into the exposed collagen network and also partially into the dentinal tubules. The interdiffusion of the synthetic adhesive polymer within the collagen network forms a micromechanical bond and is commonly referred to as a hybrid layer (Fig. 1). In the last 20 years,

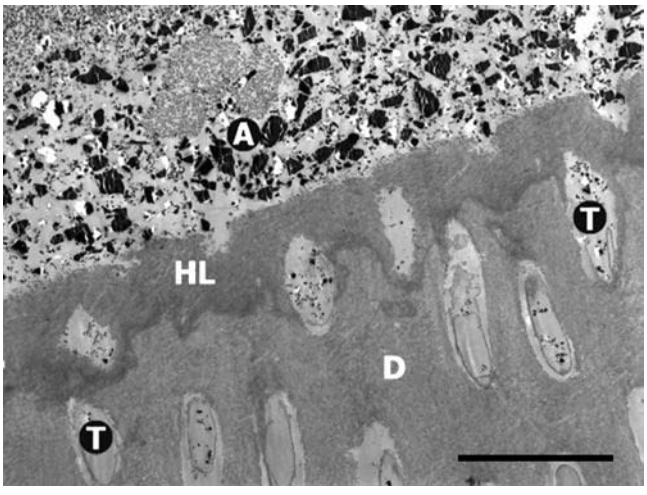


Figure 1. Scanning electron micrograph (SEM) image of a total-etch three-step dental adhesive system applied to dentin. A = filled adhesive resin (ceramic filler particles visible within resin matrix), HL = hybrid layer (interdiffusion zone of adhesive polymers and collagen network), T = dentinal tubules, D = laboratory demineralized dentin. Original magnification = 2000X, black bar=10 μm. (Photomicrograph courtesy of Marcos Vargas, University of Iowa College of Dentistry.)

dentin bonding has improved the clinical success of adhesively placed restorations, but difficulties remain. Completely penetrating the exposed collagen matrix with the adhesive monomers can be hindered by the presence of excess solvent, dentinal fluids, or by physical blockage of the interpenetrating microchannels between collagen fibrils. Even if the adhesive fully wets the dentin surface, suboptimal polymerization may reduce bonding effectiveness. This union of restorative material to enamel and dentin is critical not only for retaining restorations in place, but also for sealing the margin from the passage of bacterial fluids, molecules, or ions. Leakage between the interface of the dental restoration and the wall of the cavity preparation has been associated with marginal discoloration, secondary caries, and pulpal pathology.

Adhesive liners or bonding agents are also used in conjunction with resin-based composite cements when adhesively cementing crowns or other fixed appliances into or onto prepared teeth. The ceramic elements or oxides of the internal surface of dental porcelain–ceramic, metals, and resin-based composite restorative materials are mechanically or chemically roughened before applying a silane coupling agent. The silane bonds to the ceramic surface with both covalent and hydrogen-bonding protecting from hydrolytic degradation while making the surface hydrophobic and organophilic for resin cement wettability and copolymerization. Self-adhesive cements include the glass ionomer and polycarboxylate cements. Zinc phosphate and zinc oxide eugenol are nonadhesive cements that act as mortar or a luting agent.

Additional uses of the above mentioned cements include (*1*) cavity liners to achieve a physical barrier to bacteria and their products and/or to provide a therapeutic effect, such as fluoride release from glass ionomer or pulpal obtundent with zinc oxide eugenol, (*2*) cavity bases to block out undercuts in cavity preparations for indirect restorations or for insulating the pulp from thermal changes, and (*3*) temporary or provisional restorations. The beneficial effects of adhesion to tooth structure and fluoride release obtained from adhesive liners and glass ionomer materials are rapidly replacing the traditional liner, base, and cement materials, that is, zinc phosphate, polycarboxylate, and zinc oxide eugenol.

PROSTHETIC RESTORATIVE MATERIALS: CROWNS, BRIDGES, DENTURES

Metals and Alloys

Noble and base metal alloys are used for (*1*) crowns and bridges with fused porcelain in esthetic areas, (*2*) inlays, onlays, crowns, and bridges without porcelain veneering in the posterior or nonaesthetic regions of the mouth, and (*3*) partial and complete removable denture bases. Base metals commonly used in dental alloys include, nickel, chromium, copper, zinc, gallium, silver, indium, and tin. Silver, a "precious" metal, is not considered a noble metal in dentistry due to its corrosion in the oral cavity. The noble metals utilized in dentistry are gold, platinum, palladium, iridium, rhodium, and ruthenium. Cold-worked or wrought noble and base alloys can be cast with or "soldered" (brazed) to cast structures as attachments or clasps to removable partial dentures. High purity gold, being soft and malleable, can be cohesively adapted in the form of gold foil into small cavity preparations by careful condensation techniques. This process develops adequate hardness and physical properties through work hardening, resulting in a clinically successful, long-lasting restoration. However, the compacted gold foil restoration is becoming increasingly uncommon due to the success of adhesively bonded tooth colored restorations and the skill and time required to properly place gold foil. Almost all fixed-dental prostheses contained a minimum of 75% gold before the dramatic increase in the price of gold after the United States separated gold from monetary standards in 1969. The increase in gold prices, and rise in palladium prices three decades later, has lead to an increased use of alternative alloys containing base metals.

The lost wax technique became common in dentistry after W.H. Taggart introduced his casting machine in 1907. A wax pattern of the desired restoration is fabricated and then invested in a ceramic material (casting investment), which is subsequently heated to burn out the wax pattern, that is "lost wax". A molten metal alloy is then cast into the resultant space previously occupied by the wax. The restoration is recovered, finished, and polished before cementing or delivering to the patient. The investment must be able to expand enough upon setting and heating to compensate for the wax and metal shrinkage if a precise fit is to be obtained.

Alloys should (*1*) produce no toxic, carcinogenic, or allergic reactions; (*2*) resist corrosion and physical changes in the oral environment; (*3*) possess physical properties, that is, strength, fusing temperature, thermal conductivity, and coefficient of thermal expansion, appropriate for the desired application; (*4*) be able to be fabricated in a technically feasible manner; and (*5*) be available and relatively inexpensive. The alloys used for metal-ceramic or commonly termed porcelain-fused-to-metal (PFM) restorations must possess a fusion temperature range that is substantially higher ($>100\,^{\circ}C$) than the ceramic firing temperature, have sufficient creep resistance at that temperature, and have the ability to form a good bond between its oxide surface and the ceramic veneer.

Wrought stainless steel alloys are used in orthodontic brackets and wires, endodontic instruments, prefabricated temporary crowns and space maintainers. In addition, wrought cobalt–chromium–nickel, nickel–titanium, copper–nickel–titanium, and beta-titanium alloys are also used as orthodontic wires. Nickel–titanium and copper–nickel–titanium orthodontic wires have a unique superelastic (pseuedoelastic) property that delivers a constant low-level force over an extended range of deformation. Nickel–titanium and copper–nickel–titanium alloys are also the shape memory alloy (SMA), that is, they can be deformed plastically below its transition temperature range (TTR), then after heating through and above the TTR, they will return to their original desired shape due to a crystallographic transformation from martensitic phase into austenitic phase. Titanium and titanium alloys, especially due to their thin stable oxide layers, are very important endosseous dental implant materials and, with the

recent refinement in casting techniques, can be used for crowns, partial dentures and complete denture bases.

Ceramics

The first porcelain was developed in the T'ang Dynasty from 618 to 906 AD and the first suggested use of porcelain for dentistry was by Pierre Fauchard in France after the porcelain formula was brought from China by a Jesuit priest. Several developments followed but the current approaches to ceramic-metal crowns and bridges occurred from the 1950s–1960s. High-fusing alloys combined with the development of low-fusing thermally compatible leucitic porcelains permitted the fabrication of ceramic-metal restorations. High expansion leucite was mixed with feldspar glass during manufacturing to refine the coefficient of thermal expansion (α) creating a successful junction between dental porcelain and metal. The combination of which must be thermally compatible for fabrication and so that the veneering materials surface is left in a residual state of compression. The α of the veneering material is generally $\sim 0.5–1.0 \times 10^{-6}\,^{\circ}C^{-1}$ lower than the core material so that upon cooling the inner core will contract more resulting in a residual compressive state resisting crack formation and propagation of the relatively brittle veneering material.

Minor refinements have continued with metal–ceramic systems, but more recently significant advances have occurred in the area of "all-ceramic" systems, in which the metal is replaced with a ceramic core upon which veneering porcelain is fused. This eliminates the masking of the metal with opaquing agents and greatly simplifies the aesthetic technical procedures by the dental technician. However, more tooth structure may need to be removed in preparation for the construction of these all-ceramic restorations and their brittle nature does not yet permit their function in long-span bridges.

Ceramic systems can be classified by their fusion temperature, clinical usage, processing methods, or crystalline phase. Table 2 lists some crystalline types used for the three major applications of ceramics in dentistry: (*1*) metal–ceramic crowns and fixed-partial dentures (bridges); (*2*) all-ceramic crowns, inlays, onlays, veneers, and shortspan bridges; and (*3*) denture teeth.

Table 2. Classification of Dental Ceramic Materials

	Fabrication	Crystalline (dispersed) Phase
Ceramic–Metal (PFM)	Sintered	Leucite ($KAlSi_2O_6$)
All-Ceramic	Machined	Alumina (Al_2O_3)
		Feldspar ($KAlSi_3O_8$)
		Mica ($KMg_{2.5}Si_4O_{10}F_2$)
	Slip-cast	Alumina (Al_2O_3)
		Spinel ($MgAl_2O_4$)
		Zirconia (ZrO_2)
	Heat-pressed	Leucite ($KAlSi_2O_6$)
		Lithium disilicate ($Li_2Si_2O_5$)
	Sintered	Alumina (Al_2O_3)
		Leucite ($KAlSi_2O_6$)
Denture Teeth		Feldspar ($KAlSi_3O_8$)

[a]Adapted from Ref. (1) p. 553.

"Porcelains" are composed of kaolin (clay), feldspar, and quartz (flint), while the dental porcelains being quite similar, are fabricated from silica (SiO_2), soda (NaO_2), potash (K_2O), alumina (Al_2O_3), with the addition of pigments, opacifiers and fluxes. Naturally occurring minerals such as feldspar ($K_2O\ Al_2O_3\ 6SiO_2$), quartz, and nepheline syenite have been utilized to provide these constituents. The use of feldspar has lead to the term feldspathic porcelain, however, feldspar is not necessarily present in the final processed porcelain, nor is it essential to form leucite, the major crystalline phase of the porcelain. Like all dental ceramics, dental porcelain is composed of a glassy (vitreous) matrix phase surrounding a dispersed crystalline phase. The glassy matrix, composing 75–85% of the porcelain, is formed by heating the raw materials into a glassy state then quenching. This pyrochemical reaction produces a supercooled liquid of metastable equilibrium that is quenched then ground into a fine powder. These fine powders or frit can be reheated and will fuse at a lower temperature with little pyroplastic flow giving increased homogeneity, translucency, smoother texture, a lower fusion temperature, and less shrinkage. The temperature at which the surface glassy phase softens allowing the fritted particles to coalesce without further pyrochemical change is called sintering. These sintered dental porcelains, in general, will have little change in the physical, chemical, or optical properties of the glassy matrix upon repeated firings during the necessary steps of the restoration fabrication. However, if improperly fired or over fired, the dispersed leucite crystals can be altered leading to reduced strength or porcelain–metal (core) thermal incompatibilities. During the fritting process the silica matrix is disordered due to the rapid cooling from the molten state and also by the addition of fluxes that break up the silica tetrahedral network by occupying oxygen. These alkali ions reduce the number of cross-linkages between the silicon–oxygen tetrahedra by randomly occupying space in the open network. The net effect of flux (LiO_2, Na_2O, K_2O, BaO, CaO, MgO, ZnO) addition is lower softening or fusion temperature, decreased viscosity, production of glassy phase, increased α, decreased strength, lowered chemical resistance, increased risk of devitrification during repeated firing cycles. The lower fusing temperatures and increased α made possible the modern day metal–ceramic systems. Three components of the porcelain to metal bond are classically described: (*1*) mechanical interlocking through good wetting of the porcelain on the roughened metal surface, (*2*) chemical physical bonding between the oxides of the porcelain and the oxides on the metal surface, and (*3*) a controlled mismatch in α leading to residual compressive forces in the porcelain (described earlier). Any of these may predominate depending on the ceramic system.

Achieving superior esthetics, in general, is simplified by having a ceramic core. However, strength, wear, fit, and longevity must be proven in controlled clinical trials. Increasing the strength of the ceramic core to perform comparably to metal substructures is approached by manipulating the crystalline phase for reinforcement.

Techniques for fabricating all-ceramic systems include (*1*) sintering with alumina-based, magnesia- based, and leucite-reinforced ceramics; (*2*) heat-pressed techniques with leucite-reinforced and lithium-disilicate-based ceramics; (*3*) slip-casting with alumina-, spinel-, and zirconia-based ceramics; and (*4*) the machining of manufactured ceramic blocks available in several types of ceramic. One method uses computer-aided designing/computer aided machining (CAD/CAM) technology to fabricate inlays, onlays, veneers, and crowns. An "optical" impression is obtained from the prepared tooth with an optical scanner and the computerized image of the restoration is designed by the computer software. Subsequently, a ceramic block is machined in the computer-controlled milling machine according to the design and later cemented into or on the tooth by the dentist.

Slip-cast all-ceramics are fabricated by a process very similar to that used for the production of common objects such as plumbing fixtures and beer steins. Successive layers of ceramic slurry are applied to porous refractory gypsum that draws in the water depositing a solid layer of alumina on its surface. The ceramic buildup is dried then sintered for 4 h at 1100 °C, the porous alumina coping is then carved into the desired shape before infiltrating with a slurry of lanthanum aluminosilicate glass by firing at 1120–1150 °C for 3–5 h. The resultant ceramic is a three-dimensional (3D) interpenetrating network of alumina and glass of high strength due to the presence of densely packed alumina and low porosity. The excess glass is removed and the core is subsequently veneered with a thermally compatible veneering ceramic. Improved translucency (esthetics) can be obtained by glass infiltrating a core composed of magnesium spinel and alumina. The strongest slip-cast material currently available contains tetragonal zirconia along with alumina and glass. When a load is induced on the tetragonal zirconia it absorbs energy by transforming into a monoclinic crystal form accompanied by a volume increase of 3% in a crack arresting manner. Flexural strengths (380–700 MPa) and fracture toughness (2–7 $MPa \cdot m^{1/2}$) for these core materials are in the following rank order: spinel < alumina < zirconia.

Prosthetic Resin Materials

Poly(methylmethacrylate) was introduced as a denture base material in 1937 and in roughly a decade had virtually replaced the use of vulcanite. Acrylic polymers also enjoy a wide variety of uses in additional prosthetic applications, such as, artificial denture teeth, provisional restorations and temporary crowns, denture base repair, relining and rebasing materials, and obturators for maxillofacial defects.

Denture base materials are typically fabricated from heat-cured poly(methylmethacrylate) and rubber-reinforced poly(methylmethacrylate) and perform surprisingly well. These plastics are supplied in a powder liquid or gel form. The 10 poly(methylmethacrylate) powder is modified with ethyl, butyl, or other alkyl methacrylates for impact resistance and contains benzoyl peroxide or diisobutylazonitrile to initiate polymerization when mixed with the liquid monomer. Pigments are added to obtain natural tissue appearance, for example, mercuric sulfide, cadmium sulfide, cadmium selenide, ferric oxide, or carbon black. Various glasses, ceramics and polymer fibers have been added as dispersed phases to various products in an attempt to reinforce the acrylic polymers. The liquid component is methyl methacrylate, modified with various other monomers while including an inhibitor such as hydroquinone to prevent premature polymerization for adequate shelf life. The liquid of cold-, self-, or autocuring resins contain tertiary amine or sulfinic acid chemical accelerators to allow the polymerization of the monomer at room temperature. Plasticizers for resilience and cross-linkers for hardness and decreased solubility may also be included. Denture base resins can also be fabricated through pressure, heat and light-activated techniques with compositional modifications for the various initiation reactions and physical handling properties during fabrication. A number of general requirements for denture base resins are outlined in ANSI/ADA Specification No. 12 (ISO 1567) providing guidance to dentists and dental manufacturers.

Denture teeth are also fabricated from acrylic and modified acrylic materials and are generally preferred over porcelain denture teeth due to wear characteristics, phonetics and technical considerations during fabrication and repair. Temporary or provisional restorations are also fabricated from acrylic-based resins, placed during an interim period in or over the coronal aspects of the tooth while a crown or bridge is fabricated in the dental laboratory. Due to ease of fabrication and tooth-like appearance these are much more popular than aluminum shells or polycarbonate crowns that typically must be relined before temporary cementation to the tooth.

Defects of the head and neck resulting from cancer surgery, accidents and congenital deformities have been corrected with a wide variety of maxillofacial resin materials, including poly(methylmethacrylate), plasticized polyvinylchloride, polyurethane, heat- vulcanized and room temperature-vulcanized (RTV) silicone and a whole host of various other elastomers. It is important to use prosthetic resin materials with color stability, ease of fabrication, dimensional stability, edge strength, flexibility, low thermal conductivity, biocompatibility, and surface texture to achieve clinical success and patient acceptance. Silicones are the most widely used materials for facial restorations in the United States, with RTV Silicone MDX-4-4210, possessing surface texture and hardness within the range of human skin. The prosthesis can be held in place by tissue undercuts, the patient's glasses or dentures, medical grade adhesives, magnetic attachment to endosseous implant-retained metallic attachments or bars or through a combination of methods. A mold is made from an impression of the defect upon which a prosthesis is fabricated and color matched by mixing small amounts of pigments into the elastomer. Surface coloration and texturing is completed and the patient returns periodically for esthetic touchups to achieve a lifelike match to the skin.

Implants

The surgical placement of endosseous dental implants to support dental restorations has become a routine aspect of

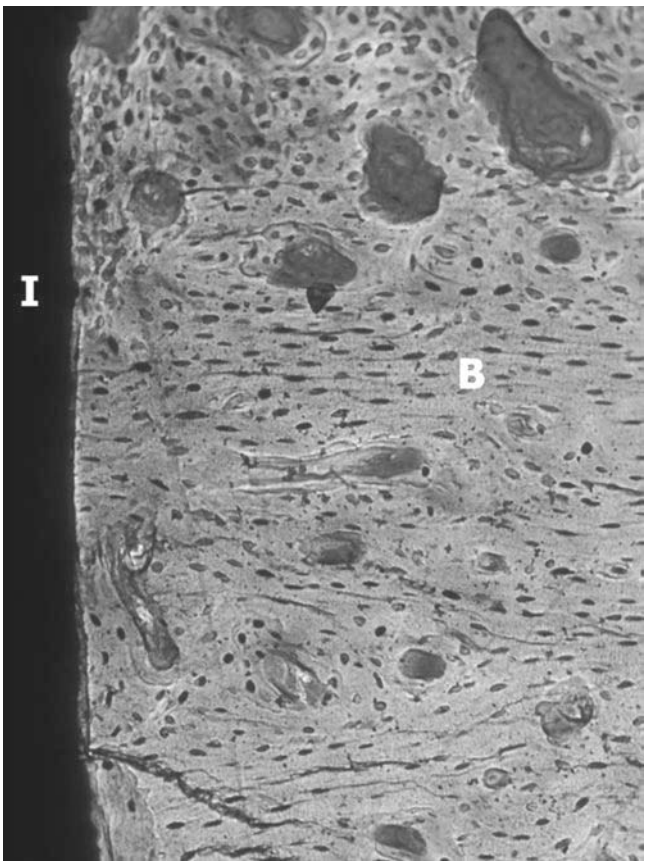


Figure 2. Dental implant demonstrating osseointegration. I = implant, B = bone. (Courtesy of John Keller, University of Iowa College of Dentistry.)

dental care enjoying a high success rate. Commercially pure titanium (CpTi) and Ti-6Al-4V are the materials most commonly used for endosseous dental implants. The stable oxides surfaces formed on CpTi and Ti- 6Al-4V have proven to successfully biointegrate with bone. The terms osseointegration and functional ankylosis are used to describe the direct bone apposition on the implant surface giving evidence to support a direct biochemical bonding (Fig. 2). The clinical success of these implants depends not only on osseointegration, but also the quantity, quality and distribution of bone at the implant site, the technical skill during surgical placement, and the timing and degree of mechanical loading under function. Many other factors play a role but six biological and technical factors are recognized as key to implant success: (*1*) implant surface texture, (*2*) biocompatibility, (*3*) implant design, (*4*) host tissue condition, (*5*) surgical technique, and (*6*) loading conditions.

Lower success is observed in areas of the mouth that may have less cancellous bone or thin cortical plates such as the posterior regions of the maxilla. Therefore, attempts are made to manipulate the osseous response to the implant so that the bone quantity and quality at the implant interface is optimized for the clinical requirements. This will not only help the routine implant but is especially important for those patients with: (*1*) poor bone quality, (*2*) heavy masticatory loading, and (*3*) the need for multiple tooth replacement. Surgical procedures are also utilized that enhance the osseous tissues at the intended implant site by auto- or allo-grafting.

Implant surface modifications have been heavily investigated with every major manufacturer offering various implant designs and surface textures. Two of the most thoroughly investigated and successful surfaces are machined titanium and titanium plasma-sprayed (TPS) surface, with the latter significantly increasing the surface area for bone contact. These rougher surfaces have been shown to require higher forces to be removed from the bone than do smoother surface implants and may allow: shorter healing periods, the use of less invasive shorter implants and may not require bicortical implant engagement. Improving the bone adaptation through microretentive mechanisms can be divided into those that attempt to enhance the immigration of new bone through surface topography, that is, osteoconduction, and those that attempt to manipulate the type of cell response and growth for new bone formation, that is, osteoinduction. Osteinductive methods also include the use of the implant as a delivery device for biomolecules for the induction of the desired response. A complex cascade of molecular and cellular processes occur after the placement of the implant into a surgical site, many of which are just now beginning to be understood, leading to the possibility of implant-mediated tissue engineering. Calcium phosphates or "hydroxyapatite" can also be coated on titanium implants and have been documented to create a very intimate bone-to-implant contact with a reduced healing period; however, the long-term results are less favorable than that achieved with TPS due to surface degradation and coating separation problems.

Macroretentive features are also part of the implant design including: screw threads, solid body press-fit designs, and sintered bead technology. These macroretentive features are intended to improve initial implant stability and enhanced bone ingrowth. Without the aid of a periodontal ligament (present between the natural tooth root and bone) the bone responds most favorable to compressive loading, which must accounted for in the implant design.

The original guidelines for implant success have changed over time with an increased use of nonsubmerged (not covered by the gum tissue) and single-staged surgical techniques (immediate abutment placement). These time- and cost-saving changes have come about after studies revealed similar end-results, in terms of "biological width" (composed of junctional epithelial and connective tissue attachment) and clinical survival rates. Considerations such as these are blurring the distinction between the clinical healing period (Phase I) and the functional period (Phase II). Additionally, studies have shown similar clinical predictability using both solely implant supported and mixed tooth-implant-supported fixed partial dentures (bridges), while the use of cemented rather than screw-retained prostheses have reduced technique complications.

Abutments of either titanium or alumina, as compared with those abutments of more "esthetic" quality, such as gold alloys or dental porcelains, are most likely to have favorable soft-tissue healing with formation of a

physiological epithelial and connective tissue attachment without subsequent bone resorption. Therefore, surgical techniques to help hide the prosthesis-abutment in the casual viewing region of a patient's mouth by careful peri-implant soft tissue manipulation with proven implant materials is currently recommended.

As work continues to optimize the osteoconductive (passive) response to implant surfaces, research will also progress toward predictable osteoinductive (active) responses. As hard and soft tissue responses are optimized through surgical protocols and biomaterial influences, healing phases will be shortened, retention rates will be increased, and loading capabilities will be improved, allowing the placement of fewer, less invasive and less expensive implants for predictable long-term, implant supported prostheses.

SUMMARY

This article briefly reviewed those commonly used metal, ceramic, polymer, and composite materials used in dentistry for the restoration of individual teeth or the replacement of missing teeth. Restorative materials include noble and base metals, resin based composites, glass ionomers, ceramics, acrylics, and amalgam alloys. These materials are either directly placed into the prepared tooth cavity or cemented in place after laboratory fabrication. An increasing number of "fillings" are retained by the use of dental adhesives. The dentist, in consult with the patient, must take several factors into consideration in the selection of restorative materials, to include (1) chewing forces, (2) esthetic demands' (3) strength of remaining tooth structure, (4) diet, (5) hygiene, and (6) cost. No one material type possesses all the desired physical properties; therefore, several materials are required for successful dental restoration.

Our population is living longer while retaining more of their teeth. With the increased emphasis on preventive dental care, increased awareness and the desire for health, our population will require more partial and single tooth restorations or replacements, especially in the area of root caries and less of a need for removable partial dentures, complete dentures and fixed bridges. Improvements in adhesive dental procedures and direct placed tooth colored resin-based composites will allow more conservative dental care. The interplay of biomaterials and biomolecules may also lead to the predictable regeneration of hard and soft tissues, while tissue engineering may someday lead to the induction of whole tooth regeneration.

BIBLIOGRAPHY

Cited References

1. Craig RG, Powers JM. Restorative Dental Materials. 11th ed. St. Louis: Mosby; 2002.

Reading List

Anusavice KJ. Phillip's Science of Dental Materials. 10th ed. Philadelphia: W. B. Saunders; 1996.

Denry IL. Recent advances in ceramics for dentistry. Crit Rev Oral Biol Med 1996;7(2):134–143.

Ferracane JL. New polymer resins for dental restoratives. Oper Dent 2001; Suppl 6: 199–209.

Keller JC. Dental Implants: The relationship of materials characterization to biologic properties. In: Bronzino JD, editor. The Biomedical Engineering Handbook. 2nd ed. Boca Raton, FL: CRC Press LLC; 2000.

Osborne JW. Mercury, its impact on the environment and its biocompatibility. Oper Dent 2001; Suppl 6:87–104.

Salvi GE, Lang NP. Changing paradigms in implant dentistry. Crit Rev Oral Biol Med 2001;12(3):262–272.

Smith DC. A new dental cement. Br Dent J 1968;125:381–384.

Stanford CM. Surface modifications of implants. Oral Maxillofacial Surg Clin N Am 2002;14:39–51.

See also BIOCOMPATIBILITY OF MATERIALS; BONE AND TEETH, PROPERTIES OF; RESIN-BASED COMPOSITES; TOOTH AND JAW, BIOMECHANICS OF.

BIOMATERIALS, POLYMERS

MIN ZHANG
University of Washington
Seattle, Washington

SUSAN P. JAMES
Colorado State University
Fort Collins, Colorado

INTRODUCTION

Biomaterials are materials of synthetic as well as of natural origin in contact with tissue or biological fluids, including metals, ceramics, polymers, and composites. The main advantages of polymeric biomaterials over ceramics and metals are the variety of composition, properties and available forms (solid, hydrogel, and solution), and ease of fabrication into complex shapes (films, sheets, fibers, powders, etc.) and structures because synthetic polymers are easily tailored to specific applications. In addition, polymeric materials are much lighter than metals and ceramics. Since most natural biomaterials are polymeric, mimicking the function and/or structure of natural materials (e.g., skin) is more easily achieved with polymers or polymeric composites than metals and ceramics. As with other biomaterials, there are some basic requirements for polymeric biomaterials. They must be (*1*) nontoxic, for example, not causing carcinogenesis, pyrogenicity, hemolysis, sustained inflammation, and allergy; (*2*) biocompatible, that is, not causing foreign body reactions, such as complement activation, thrombus formation, collagenous tissue encapsulation, calcification, and compatibility with the contact tissue in physical and mechanical properties; (*3*) sterilizable with autoclave, dry heating, ethylene oxide, gas plasma or γ irradiation, or be produced in a sterile fashion so no postmanufacture sterilization is required (1a).

Synthetic polymers have been widely used in biomedical devices, for example, hard and soft tissue implants, extracorporeal devices, drug delivery systems, and medical disposable supplies. They exhibit diverse properties, ranging from hydrophobic, non-water-absorbing materials (e.g., polyethylene, polypropylene, and polytetrafluoroethylene), to hydrophilic, water-swelling hydrogels [e.g., poly(hydroxyethyl methacrylate)], and to water-soluble materials [e.g.,

poly(vinyl alcohol)] (2a). Biological polymers are obtained from animals, plants, bacteria, or other living creatures. Their remarkable advantages over synthetic polymers include their excellent physiological activities. These activities include cellular activity regulation (e.g., hyaluronic acid), selective cell adhesion (e.g., collagen), and similar properties to natural tissues (3). Most biopolymers are biodegradable, so they are suitable for use in temporary medical devices, drug-delivery systems, and tissue engineering scaffolds.

This section introduces the synthesis methods of generic polymers, and the effect of composition and structure on their properties. Following this, 13 groups of polymers that have found wide biomedical application are reviewed and their properties and uses are discussed.

POLYMERIZATION–SYNTHESIS

Polymers are long molecules made up of a large number of simple repeating units. They are prepared from monomers through a process called polymerization. Small monomer molecules react chemically to form either linear chains or three-dimensional (3D) networks. Conventionally polymerization mechanisms are divided into two main categories: condensation polymerization (also called step-growth polymerization) and addition polymerization (also called chain polymerization).

Condensation or Step-Growth Polymerization

Condensation polymerization occurs between an organic base (e.g., an alcohol and amine) and an organic acid (e.g., carboxylic acid and acid chloride), and a small molecule (e.g., water) is condensed out during the reaction. In a condensation reaction to combine two monomers together to form a dimer, each monomer molecule loses an atom or a group of atoms at the reactive end, leading to the formation of a covalent bond between the two molecules, while the eliminated atoms bond with others to form small molecules. The dimers can react with each other or with unreacted monomers until finally a long molecule is generated. An equilibrium exists between the reactants and products during condensation polymerization. The condensate (e.g., water) should be removed to drive the reaction toward the product direction. A high molecular weight product can be obtained only after a sufficiently long reaction time.

The reaction between a diacid and a dialcohol to produce an ester is a typical condensation polymerization example.

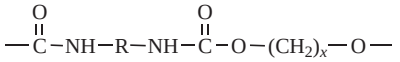
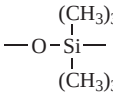
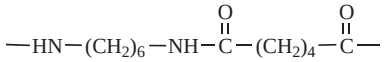
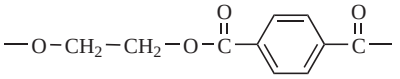
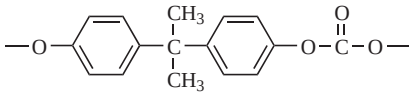
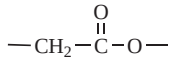
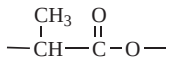
$$n\,\mathrm{HOOC{-}R_1{-}COOH} + n\,\mathrm{HO{-}R_2{-}OH} \rightarrow$$

$$\mathrm{HO}\left[\mathrm{{-}OCR_1COOR_2{-}O}\right]_n\mathrm{H} + (2n-1)\mathrm{H_2O}$$

Some condensation polymers used as biomaterials are given in Table 1. They are typically synthesized from reactions of acids and alcohols to produce polyesters, reactions of acids with amines to produce polyamides, or reactions of alcohols or amines with isocyanates to produce polyurethanes or polyurea, respectively.

Synthesis of biopolymers is very complicated, but typically involves enzyme-catalyzed condensation polymerization occurring in animal or plant cells, or in microorganisms via their metabolic pathways (4,5a).

Table 1. Typical Condensation Polymers

Polymer	Repeat Unit
Polyurethane	
Silicone rubber	
Polyamide (Nylon 66)	
Poly(ethylene terephthalate)	
Polycarbonate	
Polyacetal	
Polyglycolic acid (PGA)	
Polylactic acid (PLA)	

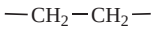
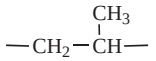
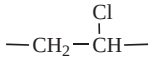
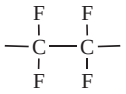
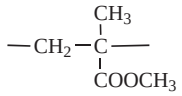
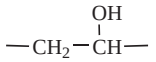
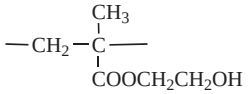
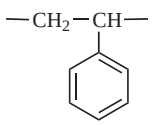
Addition or Chain Polymerization

Addition polymerization occurs among small molecules with double bonds. Polymer chains are formed by opening-up double bonds of unsaturated monomer units and successive addition to a growing chain with an active center. No small molecule byproducts are formed during addition polymerization. Consequently, the composition of the repeating unit of the polymer is identical to that of its monomer. Addition polymerization takes place in three distinct steps: initiation, propagation, and termination. Initiation occurs by an attack on the monomer molecule by a free radical, a cation, an anion, or Ziegler–Natta catalysts; accordingly, addition polymerization can be divided into four types: free-radical polymerization, cationic polymerization, anionic polymerization, and coordination polymerization. No matter how the reaction is initiated, once a reactive center is created, many monomers are added onto it and the molecule chain grows very large within a few seconds or less, so the addition polymer size (i.e., molecular weight) is independent of reaction time. Unlike condensation polymerization, no dimer, trimer, or other intermediates can be found in addition polymerization. Polymerization of ethylene is a typical example of addition polymerization.

$$n\,\mathrm{H_2CCH_2} \rightarrow \left(\mathrm{H_2C{-}CH_2}\right)_n$$

Typical addition polymers are listed in Table 2.

Table 2. Typical Addition Polymers

Polymer	Repeat Unit
Polyethylene	
Polypropylene	
Polyvinyl chloride	
Poly(terafluoroethylene) (Teflon)	
Poly(methyl methacrylate)	
Poly(vinyl alcohol)	
Poly(hydroxyethyl methacrylate)	
Polystyrene	

The stereoregularity and branching of addition polymers can be controlled through varying the type of initiator and the reaction conditions (6). Inonic addition polymerization can lead to some control of tacticity and a stereoregular structure. Polymers produced through coordination polymerization have a high degree of stereoregularity. The polyethylene produced with peroxide initiator is highly branched. By using a Ziegler–Natta catalyst, linear, high density polyethylene or ultrahigh molecular weight polyethylene (UHMWPE) can be obtained. Living free-radical polymerization is a newer form of free-radical polymerization. By using rapid initiation, slow propagation, and inhibition of termination and transfer reactions, the molecular structure of polymers can be precisely controlled. This method can be applied to vinyl monomers to produce block, graft, star polymers, polymer brushes, and many other architectures (5a,7).

Molecular Weight and Its Distribution

Almost all polymers consist of molecules (a.k.a., chains) with a variety of lengths, so it is only possible to quote an average value of molecular weight. The length of a polymer molecule is represented by the degree of polymerization (DP), which is equal to the number of repeat units in the chain. The relationship between polymer molecular weight (MW) and degree of polymerization can be expressed as

$$\text{MW of polmer} = \text{DP} \times \text{MW of repeating units}$$

The number-average molecular weight (M_n) and weight-average molecular weight (M_w) are the most commonly used average values of molecular weight. The number average molecular weight is defined as the sum of the products of the molecular weight of each fraction (M_i) multiplied by its mole fraction (x_i) (Eq. 1), which can be obtained using gel filtration chromatography, light scattering, or ultracentrifugation. Whereas M_w is the sum of the products of the MW of each fraction (M_i) multiplied by its weight fraction (w_i) (Eq. 2), which can be measured with osmometry.

$$M_n = \sum x_i M_i \tag{1}$$

$$M_w = \sum w_i M_i \tag{2}$$

The ratio of M_w/M_n is defined as the polydispersity index (PDI), representing the breadth of the molecular weight distribution. When all the polymer chains have the same length, the ratio is 1. A low polydispersity index is necessary to control physical and mechanical properties of polymers because the short chains usually present when PDI is high degrade properties.

COMPOSITION, STRUCTURE, AND PROPERTIES

The structure and behavior of polymers is strongly temperature dependant. There are two major transition temperatures for polymers: T_g and T_m. The glass-transition temperature is a second-order transition temperature, associated with the amorphous regions of polymers. It marks the onset of significant molecular motion. Above T_g, polymers soften, and become rubber-like and more easily deformed. Below T_g, polymers become hard and brittle, and glass-like. Applications of polymeric biomaterials are related with their T_g values. For example, silicone rubber with a T_g of $-127\,^{\circ}C$ is soft and acts as an elastomer at $37\,^{\circ}C$ (human body temperature), while poly(methyl methacrylate) (PMM) used as bone cement with a T_g of $105\,^{\circ}C$ retains high strength, stiffness, and creep resistance at $37\,^{\circ}C$. The melting temperature is associated with crystalline regions of polymers. It is a first-order transition temperature. Above T_m, the polymer is in a melt liquid state. Polymeric materials with a low T_m value can be melt processed.

Polymer molecules can be linear, branched, or a cross-linked network. Schematic representations are given in Fig. 1. Based on their molecular structure and their mechanical and thermal behavior, polymers are classified into three major categories: thermoplastics, thermosets and elastomers. They are discussed in detail respectively.

Thermoplastics

These polymers soften and harden reversibly with changes in temperature. Both linear and branched polymers are thermoplastic. The properties of thermoplastics can be changed by controlling the following factors.

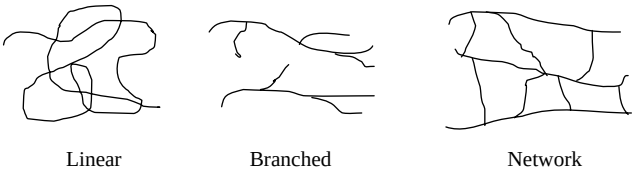
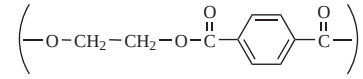


Figure 1. Schematic representation of different types of polymer molecules.

Molecular Weight. The molecular weight and its distribution have a great effect on the properties of thermoplastics. By increasing molecular weight, the polymer chains become longer and more entangled, resulting in a higher melting temperature and improved strength, including resistance to creep (8a). The strength properties increase with the molecular weight rapidly at first, but level off after reaching a certain point. Uniform molecular length is also important, because short molecules act as plasticizers, which decrease the mechanical properties of polymers. For example, ultrahigh molecular weight polyethylene must have a high molecular weight (MW, $2–4 \times 10^6$ g·mol^{-1}) and narrow MW distribution (i.e., low PDI) to obtain excellent mechanical properties for orthopedic applications.

Chemical Composition. The changes in the composition of the backbone or side chains also affect the properties of polymers. When atoms that can increase the flexibility of polymer chains (e.g., O and S) are incorporated into the carbon backbone, for example, polyethylene oxide (CH_2CH_2O) (T_g: $-41\,^\circ$C; T_m: $69\,^\circ$C), the glass transition and melting temperatures will decrease. On the other hand, the insertion of groups that stiffen the polymer chains markedly raises the T_g and T_m and leads to higher strength and stiffness, as seen in the case of polyethylene terephthalate  (T_g: 60/85 °C; T_m: 280 °C).

The replacement of pendant hydrogen atoms in polyethylene by other atoms also changes the polymer properties. Large atoms or groups (e.g., methyl groups in polypropylene), hinder the rotation about the backbone, resulting in higher T_g and T_m, strength, and stiffness. Similar results are observed for the polymers with polar pendant atoms or groups [e.g. poly(vinyl chloride), PVC]. van der Waals forces are enforced and even hydrogen bonds may be formed among the chains of these polymers.

Branching. Branching prevents dense packing and crystallization of the polymer chain, and thus reduces the density, melting temperature, strength, and stiffness of polymers. For example, branched low density polyethylene (LDPE) is much weaker than linear high density polyethylene (HDPE).

Tacticity. When the repeat units of a polymer are nonsymmetrical, the location of the side atoms or groups also plays an important role in the structure and properties of the polymer. If all the side atoms or groups are on one side of the main chain, the polymer is termed isotactic. In a syndiotactic polymer, the side groups alternatively appear on both sides of the chain. In both cases, the polymer is stereoregular or stereotactic. Polymer chains with stereoregularity are better able to crystallize, resulting in high T_m, stiffness and less solubility. In an atactic polymer, side groups are randomly distributed along the molecular chain. Atactic polymers give poor packing, low density, low stiffness, and strength. A typical example of the importance of tacticity is polypropylene. Isotactic and syndiotactic polypropylene have a high T_m (176 °C for isotactic, 150 °C for syndiotactic) and good mechanical properties. They are widely used as biomaterials, whereas atactic propylene is an amorphous waxlike polymer without any application in biomedical field (8b,9).

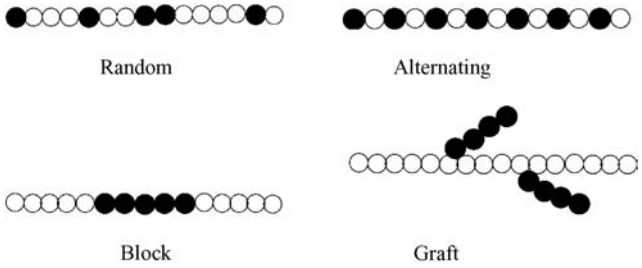


Figure 2. Four types of copolymers.

Copolymer. Copolymers contain two or more different types of repeat units. According to the distribution of the repeat units, they are divided into four types (Fig. 2). Copolymers are synthesized to obtain the desirable combination of properties of simple homopolymers.

Temperature and Time. Amorphous thermoplastics exhibit viscoelastic behavior, meaning that their properties are time or temperature dependent. At low temperatures or high rates of loading, the polymers behave in a brittle manner. However, at high temperature or low rates, the materials behave as a viscous liquid with chains easily passing one another. The application temperature of most biomaterials is 37 °C, which cannot be changed, but the properties of the materials can be controlled by selecting appropriate T_g.

Thermosets

Thermosets are highly cross-linked 3D molecular networks. High density cross-links between molecules restrict the motion of the chains of thermosets, leading to a high T_g, good strength, stiffness, and hardness, but poor ductility. The hard and stiff thermosetting polymers find uses in hard tissues (e.g., bone and teeth). Poly(carboxylic acid) cross-linked with zinc is a hard and rigid cement used for dental restorations (2a,10a). Epoxy resin is sometime used to fill the cavities of the teeth and provide hardness.

Elastomers

Elastomers are lightly cross-linked macromolecular networks. The cross-linking can be covalent or physical. In

thermoplastic elastomers, the hard and tightly packed domains with high T_g (e.g., the hard isocyanate segments in polyurethane elastomer) act as physical cross-links. The loose cross-links prevent viscous plastic deformation while retaining large elastic deformation, so elastomers can be easily stretched to high extension and will go back to their original position on removal of the stress. To act as a biomedical elastomer, a T_g much lower than 37 °C is required, and the polymer should not easily crystallize.

POLYMERS USED AS BIOMATERIALS

Many polymers have been synthesized for biomedical applications. This section just focuses on 13 groups of polymers most commonly used in clinical practice. Each part of the following deals with one polymer or one group of polymers with similar structures. The synthesis, structure, properties, and applications of these polymers will be discussed. Their trade names and related ASTM standards are listed in Tables 3 and 4, respectively, while the thermal and mechanical properties are summarized in Table 5.

Polyolefins

Polyolefins are a group of thermoplastics polymers derived from simple olefins. The most important polyolefins are polyethylene, polypropylene, and their copolymers.

Polyethylene. Commercially available polyethylene has four major grades: LDPE, linear low density (LLDPE), high density (HDPE), and UHMWPE. These materials have good toughness and excellent chemical resistance, and can be easily processed into products at low cost.

Low density polyethylene is produced through the free-radical polymerization of ethylene gas at high pressure (100–300 MPa) in the presence of peroxide initiator (5c). The synthesis conditions lead to the highly branched structure, low density (0.915–0.935 $g{\cdot}cm^{-3}$) and crystallinity of LDPE (11a). By using Ziegler–Natta catalyst, HDPE can be synthesized at a low temperature (60–80 °C) and pressure (~10 MPa). Unlike LDPE, HDPE is linear. The linearity leads to good packing of the molecular chains, high crystallinity, and density (0.94–0.965 $g{\cdot}cm^{-3}$) of HDPE. The LLDPE is produced by a low pressure process in the presence of metal catalysts. Up to 10% of a 1-alkene (e.g., butene-1, hexene-1, and octane) is used as the comonomer. Unlike LDPE, the side chains of LLDPE are very short, resulting in better properties than LDPE (12). All three types of polyethylene can be melt-processed through extrusion or molding. The LDPE cannot withstand sterilization temperatures, so only HDPE and LLDPE are used for biomedical applications (2a). The HDPE is used in tubing for catheters and drains, and in pharmaceutical bottles and nonwoven fabrics. The LLDPE is frequently used for pouches and bags due to its excellent puncture resistance. Biocompatibility tests for PE used as human tissue contact devices, short-term implantation of 30 days or less and fluid transfer devices are given in ASTM F 639 (not applicable for UHMWPE).

When molecular weights of the linear polyethylene obtained through Ziegler–Natta catalyst are $>1\times10^6 g{\cdot}mol^{-1}$, there is a sudden jump in the properties (13). The melt viscosity becomes extremely high so that the polyethylene cannot be processed with conventional extrusion and injection molding. Also it is practically insoluble in all solvents, so only sintering at high temperature and pressure may be used to fabricate the desired products. The polyethylene has excellent mechanical properties and a very low coefficient of friction and wear. It is termed UHMWPE. The UHMWPE, currently

Table 3. Structures and Trade Names of Polymeric Biomaterials

Polymeric Biomaterials	Structure	Trade Names
Cross-Linked UHMWPE	UHMWPE Cross-Linked through Radiation or Chemical Reactions	Crossfire (Stryker Howmedica Osteonics), Marathon (DePuy Orthopaedics), and Durasul (Sulzer Medica)
Polypropylene	Linear macromolecules	Prolene (Ethicon), Surgipro (Syneture)
PTFE	Linear macromolecules, highly crystalline	Teflon (DuPont)
Expanded PTFE	PTFE with microporous structure	GoreTex (Gore)
PMMA	Atactic, linear macromolecules	Plexiglas (Rohm & Haas), Lucite (DuPont)
Polyurethane	Thermoplastic segmented polyurethane elastomers	Biomer (Ethicon), Pellethane (Dow Chemical), Tecoflex (Thermedics)
Polyamide	Linear macromolecules, strong intermolecular hydrogen bonding	Nylon (DuPont)
PET	Linear, stiff macromolecules	Dacron (DuPont)
Polyacetal	Closely packed, linear molecules	Delrin (DuPont)
Poly(hydroxyethyl methacrylate)	Hydrogel	Hydron (Hydron Technologies)
PGA	Biodegradable polymer	Dexon (American Cyanamid)
PLGA	Biodegradable polymer	Vicryl (Ethicon)
Poly(ethylene oxide/propylene oxide) copolymers	PEO-PPO-PEO triblock copolymer	Pluronic F127 (BASF)
Hyaluronan	Crosslinked hyaluronan Hyaluronan with carboxyl groups esterified by alcohol	Hylan (Biomatrix)Hyaff (Fidia Advanced Biopolymer)

Table 4. Polymeric Biomaterials and Related ASTM Standards

Polymer	ASTM Standards	Scope
Polyethylene	F 639	Specifies requirements and physical/biological test methods for PE plastics used in medical devices (not applicable to UHMWPE).
UHMWPE	F 648	Specifies property requirements for UHMWPE powder and fabricated forms used for surgical implants, such as joint implants.
PVC	F 665	Classifies formulations of PVC plastics used for short-term biomedical application.
PTFE	F 754	Specifies the performance of PTFE in sheet, tube and rod shapes used for surgical implants.
PMMA bone cement	F 451	Specifies composition, physical performance, packaging requirements, and biocompatibility of acrylic bone cement.
	F 2118	Provides test methods to evaluate the fatigue properties of acrylic bone cement.
Polyurethane	F 624	Provides guide to evaluate thermoplastic polyurethane in solid and solution forms for biomedical applications.
Silicone rubbers	F 2038	Provides information about formulation and use of silicone elastomers, gels, and foams used in medical applications.
	F 2042	Provides information about fabrication and processing of silicone elastomers, gels and foams used in medical applications.
Polycarbonates	F 997	Specifies requirements and test methods for polycarbonates used for medical devices.
Polyacetal	F 1855	Specifies requirements and test methods for polyacetal used for medical devices.
L-PLA	F 1925	Specifies requirements for virgin poly(L-lactic-acid) resin used for surgical implants.
	F 1635	Defines testing methods to assess biodegradation rates and changes in material and properties of poly(L-lactic-acid) resin and devices.

used for fabrications of acetabular cups in hip replacements (Fig. 3), tibial plateau and patellar surfaces in knee replacements, sliding core in spinal disk replacements, and glenoid components in shoulder replacements, has a MW $\sim 2\text{–}4 \times 10^6$ g·mol^{-1}. The property requirements of UHMWPE for surgical implants are defined by ASTM F648. Before 1995, γ radiation in air was a standard method to sterilize UHMWPE orthopedic implants. However, free radicals within UHMWPE from gamma radiation caused oxidation and property degradation on shelf aging and *in vivo*. By 1998, all of the major orthopedic manufactures in the United States changed to use gamma radiation in an inert or a reduced oxygen environment, ethylene oxide, or gas plasma to sterilize UHMWPE (14). Despite the recognized success of UHMWPE as loading bearing surfaces in joint arthroplasties, UHMWPE wear debris and associated osteolysis and loosening of implants remains a major obstacle limiting the longevity of current joint replacements. To further improve the wear resistance, highly cross-linked UHMWPE is developed by

Table 5. Properties of Polymeric Biomaterials[a]

Polymer	T_g, °C	T_m, °C	Tensile Strength, MPa	Tensile Modulus, GPa	Elongation, %
Polyethylene					
LDPE	−120	115	7.6	0.096–0.26	150
HDPE	−120	137	23–40	0.41–1.24	400–500
UHMWPE	−120	130–145	30	1.1–2.0	300
Polypropylene	−20	175	28–36	1.1–1.55	400–900
Polyvinyl chloride	80	180	40–50	2.4–4.1	2–80
Teflon	117	327	15–35	0.40	2–5
Poly(methyl methacrylate)	105		50–75	2.0–6.0	2–10
Polyurethane (elastomer)			23–58		400–600
Silicone rubbers	−123				
Soft			6.0		600
Hard			7.0		350
Polyamide					
Nylon 6	50/100	270	70	0.7	300
Nylon 66	50	280	75		300
Poly(ethylene terephthalate)	60–85	280	50–70	3.0–4.0	30–300
Polycarbonate	150	230	65	2.4	110
Polyacetal	−85	181	65–80	5.0–13.0	9.5–12
Poly(lactic acid)					
L-PLA	54–59	159–178	28–50	1.2–3.0	2.0–6.0
DL-PLA	51		29	1.9	5.0
Poly(glycolic acid)	35	210			
Collagen fibers			50–1000	1.0	10
Cellulose acetate	230		13–60	0.45–2.8	1.9–9.0

[a]Refs. 5b 9, 10c, and 17e.

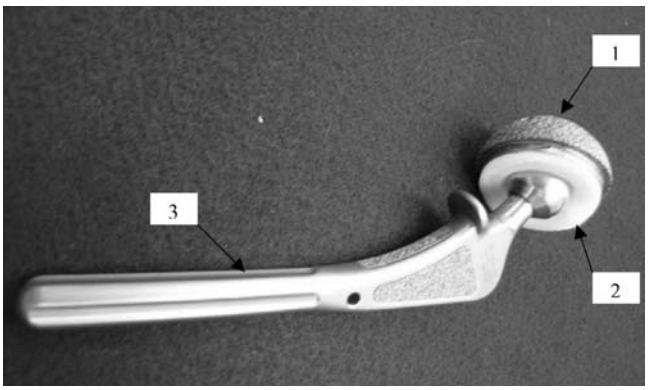


Figure 3. A total joint replacement: (1) metal backing of acetabular cup, (2) UHMWPE acetabular cup, (3) metallic femoral prosthesis.

cross-linking conventional UHMWPE through chemical reactions, or through gamma or electron beam radiation. Cross-linked UHMWPEs available on the market include Crossfire (Stryker Howmedica Osteonics), Marathon (DePuy Orthopaedics), and Durasul (Sulzer Medica). Both research and clinical applications have demonstrated that cross-linking dramatically reduces the wear rate of UHMWPE (14).

Polypropylene. Like linear polyethylene, syndiotactic and isotactic polypropylene are also polymerized through Ziegler–Natta catalyst. Although polypropylene is similar to polyethylene in structure, polypropylene has a lower density ~ 0.90 g·cm^{-3} and a higher T_g ($-12\,^\circ$C) and T_m (125–167 °C). The higher melting temperature makes polypropylene suitable for autoclave sterilization (15). The chemical resistance of polypropylene is similar to high density polyethylene, while its stress-cracking resistance and creep resistance is superior to that of polyethylene. It has an exceptionally long flex life, and thus is used to make integrally molded hinges in finger joint prostheses (11a). Also as a suture material, polypropylene yarn (e.g., Prolene from Ethicon, Surgipro from Syneture) has been used clinically (2a,10b). It causes least fibroblastic response compared with other nondegradeable suture materials and does not lose strength after it is implanted.

Poly(vinyl chloride)

Poly(vinyl chloride) is a linear, atactic polymer synthesized through free-radical polymerization. Due to the large volume and high polarity of the chlorine atoms, it is difficult for the molecular chains to rotate and disentangle and hydrogen bonds are formed between adjacent chains, resulting in high strength and stiffness, and T_g (80 °C) and T_m (180 °C) (10c). Pure PVC is hard and brittle, but with the addition of plasticizers, it becomes soft and flexible. In the medical formulations of PVC, di-2-ethylhexylphthalate (DEHP or DOP) is used as a plasticizer. Plasticized PVC is used in temporary blood storage bags, catheters, cannulae, and dialysis devices. The PVC may pose problems for long-term applications because of possible extraction of the plasticizer by body fluid. Standard classification for vinyl chloride plastics used in biomedical applications is provided by ASTM F665.

Poly(tetrafluoroethylene)

Poly(tetrafluoroethylene) (PTFE), commonly known as Teflon (DuPont), is made from tetrafluoroethylene through free-radical polymerization in the presence of excess of water for removal of heat. The polymer is highly crystalline ($>94\%$ crystallinity, T_m 327 °C), dense (2.2 g·cm^{-2}) and insoluble in all common solvents. It is very stable both thermally and chemically, and as a result it is very difficult to process. The PTFE can only be sintered into products at a temperature $>327\,^\circ$C under pressure.

The PTFE has excellent lubricity, its coefficient of friction is very low (0.1), but it is not wear resistant, its modulus of elasticity and tensile strength are very low, and even more importantly it can not maintain shape very well due to the cold-flow (5c). The use of Teflon as the acetabular component material by Charnley (2b) in his total hip replacement design 40 years ago caused a catastrophe. All Teflon cups failed *in vivo*, requiring revision surgery.

Although not suitable for load bearing surfaces, PTFE can be used for other biomedical applications because of its excellent biocompatibility and stability. Standard specifications for the implantable PTFE are given in ASTM F754. Expanded PTFE (ePTFE) vascular grafts, made by stretching paste-extruded PTFE tubes at a temperature $<T_m$ and then sintering, are soft microporous tubes (GoreTex). They show good clinical results as medium diameter (5–11 mm) vessel grafts, (e.g., femoral and popliteal artery replacements) (10d,16a). However, intimal hyperplasia of smooth muscle cells at the anastomosis frequently leads to their failure. The ePTFE grafts are also popular in hemodialysis as an interposition between radial artery and cubital vein. However, thrombosis occurring at the graft venous ends may be a concern (16b). The PTFE fabrics find applications in heart valve prosthesis as suture ring (10b). Sheets or films of PTFE or its composite with graphite are widely used by plastic surgeons in reconstruction of the maxillofacial areas (10b). The PTFE tubes are used for middle ear drain, while PTFE shunts are used to carry cerebral spinal fluid from brain to venous in the treatment of hydrocephalus.

Poly(methyl methacrylate)

The commercially important poly(methyl methacrylate) (PMMA) is atactic, which is produced by free-radical polymerization of methyl methacrylate monomer (liquid) using initiator, or thermal, or photochemical initiation. Because of the bulky side chains, atactic PMMA is completely amorphous (T_g: 105 °C) and has an excellent light transparency (92% transmission) and a high index of refraction (1.49). The transparent material is familiar as Plexiglas or Lucite. For the same side group argument presented above, the strength and stiffness of PMMA are also relatively high. As a hard thermoplastic polymer, PMMA can be easily formed into any desirable shapes by regular cast, molding, or machining.

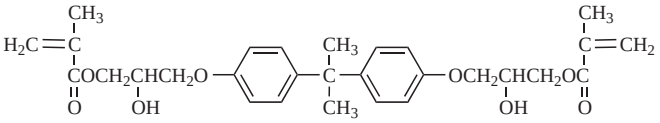


Figure 4. Chemical formula of bisphenol-A-glycidyl dimethacrylate (a new bone cement).

Poly(methyl methacrylate) is highly biocompatible and chemical resistant. It has been used in a variety of medical applications including hard contact lenses and intraocular lenses, membranes for blood dialysis, cranioplasty, bone cement for joint prostheses fixation, dentures, and maxillofacial implants.

The PMMA resin used for bone cement and dentures are usually formulated from two components: prepolymerized PMMA solid particles and the methyl methacrylate monomer liquid. When the two components are mixed during a clinical procedure, an easily moldable dough is obtained that cures in ~ 10 min. The liquid monomer polymerizes by free-radical reaction and binds the solid particles together. The composition of a commercial bone cement product (Surgical Simplex) is given in Table 6. Dibenzoyl peroxide is included in the solid component to initiate polymerization of the methyl methacrylate monomer. *N*, *N*-Dimethyl-*p*-toluidine is an activator to promote self-curing of the monomer at room temperature. Hydroquinone is added as an inhibitor to prevent premature polymerization of the monomer during storage. Barium sulfate ($BaSO_4$) is a radioopacifier. The methyl methacrylate–styrene copolymer is added to adjust the mixing and handling characteristics (e.g., viscosity, exotherm) of the cement. The physical and mechanical properties of the cement can be controlled through changing the composition and relative proportions of the components, solid particle size and its molecular weight. The requirements for PMMA bone cement are given in ASTM F451. The test methods for its fatigue performance are provided in ASTM F2118.

The PMMA bone cement is inert (no bioactivity), and the fatigue failure occurring at the cement-prosthesis and the cement-bone is a main cause of implant loosening (16c). New bioactive bone cements (BABCs) have been developed to improve the bonding strength and biochemical properties of PMMA bone cement (17a). These BABCs consist of bioactive glass ceramic powder (e.g., $CaOMgOSiO_2P_2O_5CaF_2$) and a bisphenol-A-glycidyl dimethacrylate-based resin shown in Fig. 4.

Table 6. Composition of PMMA Bone Cement (Surgical Simplex)

Components	Composition	Amount, %
Powder (40 g in a packet)	Polymethyl methacrylate	15.0 (w/o)[a]
	Methyl methacrylate –styrene copolymer	75.0 (w/o)
	$BaSO_4$	10.0 (w/o)
	Dibenzoyl peroxide	Trace
Liquid (20 mL in an ampoule)	Methyl methacrylate	97.4 (v/o)[b]
	N, *N*-Dimethyl-*p*-toluidine	2.6 (v/o)
	Hydroquinone	Trace

[a]w/o: weight percentage.
[b]v/o: volume percentage.

Polyurethanes

Polyurethanes are a family of heterogeneous polymers containing the urethane linkage (NHCOO) and frequently the urethane groups are not the predominant functional groups (5d,18). The most common method to synthesize polyurethanes consists of two steps (Fig. 5). The first step involves formation of an isocyanate terminated prepolymer from polyester or polyether polyols and di- or higher isocyanate. Subsequent reaction of the prepolymer with a chain extender, usually a diol or diamine, produces a multiblock copolymer. These two reactions are a step-growth polymerization, but no condensed byproduct is eliminated, so this type of polymerization is often referred to as a polyaddition or rearrangement polymerization. Due to the multiple choices in the chemistry and molecular weight of the various components, polyurethanes exhibit a broad range of physical properties: from hard and brittle thermoset polymers, through thermoplastic elastomers, to viscous materials. Thermoplastic segmented polyurethanes usually are valuable in producing medical devices (e.g., extruded blood tubing), while the cross-linked ones have received more attention for long-term devices and implants. The ASTM F624 provides test methods to evaluate properties of thermoplastic polyurethanes.

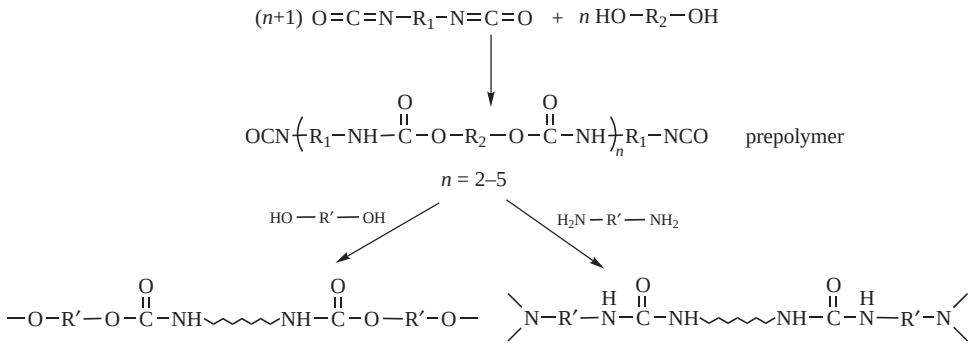


Figure 5. Two-step synthesis of polyurethane.

The isocyanates used for linear polyurethane elastomers are aromatic diisocyanates [e.g., toluene diisocyanate (TDI) and 4,4′-diphenylmethane diisocyanate (MDI)], which comprise the hard segments of the thermoplastic elastomers with the chain extenders. The soft segments of the elastomers are blocks of polyether polyols (usually polyethylene glycol PEG) or polyester polyols. These two types of segments tend to aggregate into different domains, resulting in microphase separation with the hard blocks acting as physical crosslinks in the thermoplastic elastomer.

Polyurethane elastomers are good materials for use in medical devices due to their good mechanical properties and blood compatibility. The very high flexural endurance of polyurethanes makes them major candidates for cardiovascular implants. Biomer (Ethicon, NJ), Pellethane (Dow Chemical, TX), and Tecoflex (Thermedics, MA) are all polyurethanes under different trade names. They have been widely used for cardiac guiding catheters, pacemaker lead insulation, vascular prostheses, artificial heart assist devices, blood tubing, and hollow fiber dialysers. Polyurethanes are also used extensively for wound healing. Bioclusive (Johnson and Johnson Medical, Inc.) and Opsite (Smith and Nephew) are two types of nonabsorbent wound dressings made from polyurethane films.

Although these ether-based polyurethanes are stable *in vitro*, environmental stress cracking after implanted makes their long-term biostability questionable. The microcracks caused by biological peroxidation of the ether linkage not only weaken the materials, but also serve as nucleation sites for thrombus formation (19a). To solve the problem, polycarbonate-based polyurethanes have been developed and investigated to provide an unsurpassed combination of biostability, strength, flexibility, and ease of manufacture.

Silicone Rubbers

The basic repeat unit of silicone rubbers is dimethyl siloxane. They are made by vulcanization of silicone prepolymers or by ring-opening polymerization of octamethylcyclotetrasiloxane. Silicone prepolymers are obtained by the hydrolysis of dimethyldichlorosilane with water. The hydrolysis product is dimethylsilanol, which is unstable and condenses to low molecular weight silicone prepolymers in the presence of hydrochloric acid (5d,12,17b) (Fig. 6). Silicone prepolymers are also useful as silicone oil. The vulcanization of the prepolymers can be performed at room or at high temperature. Room temperature vulcanization silicone rubbers are available in two formats: one-component or two-component. The one-component silicone rubbers result from a reaction between atmospheric moisture and a mixture of silicone prepolymers and catalyst, whereas the two-component systems result from a reaction between prepolymers and a cross-linker added to initiate the reaction. In heat vulcanization, peroxides are used to produce free radicals on heating that react with the side groups of prepolymers to form cross-links. Several copolymer silicone rubbers have been developed to meet diverse biomedical applications. Copolymers made from dimethyl siloxane and a small amount of methylvinyl siloxane result in medium and hard grades of silicone rubbers. Soft grades of silicone rubber are copolymers with phenyl-methyl-siloxane. Requirements for medical grade silicone gels and elastomers are provided in ASTM F2038 and F2042.

Silicone rubbers have many excellent properties, (e.g., extreme inertness, nonadhesion, high oxygen permeability, thermal and oxidative stability, and high flexibility at low temperature). A major disadvantage of silicone rubbers is poor resistance to tearing. Silicone rubbers are one of the widely used polymeric biomaterials in modern medicine. Since Alfred Swanson introduced silicone rubber for small flexible joints in 1960s, $>600{,}000$ silicone rubber hinge or end-bearing prostheses have been implanted to treat arthritic conditions of finger and wrist joints. These silicone implants are successful in relieving pain and restoring motion of joints at initial stage, but their long-term durability and biocompatibility have been questioned. Silicone microparticles from fragmentation or wear may cause immune reactions (20a). The largest use of silicone rubber and gel has been in breast augmentation and reconstruction (19b). Gel bleed, calcified deposit, and autoimmune diseases are concerns related to these gel-filled silicone rubber bag implants. In the earliest prosthetic cage-and-ball heart valves, silicone rubber was used for the ball, but uptake of blood lipids by the silicone led to the swelling and fracturing of the ball after several years of service. Presently, silicone rubber is only used for the suture ring in bioprosthetic heart valves (10d). Other medical applications of silicone rubber include soft contact lenses, catheters and drainage tubing, oxygenator membranes, wound dressings, and facial implants.

Polyamides

Polyamides are known as nylons. They are divided into two types: dyadic nylons and monadic nylons. The dyadic nylons, (e.g., nylon 66 and nylon 610) are made through condensation polymerization of diamine and dicarboxylic acid or its derivatives (Fig. 7a). There are two numbers following the name, the first for the number of carbon atoms in the diamine and the second for that in the diacid. The monadic nylons (e.g., nylon 6 and nylon 11) are made through self-amidation of an amino acid or through ring-opening polymerization of a cyclic lactam (Fig. 7b). The single number following the name represents the number of monomer carbon atoms. These polymers have high crystallinity and very strong intermolecular hydrogen bonding between amide groups. Thus, they have excellent fiber forming properties, and the strength along the fiber

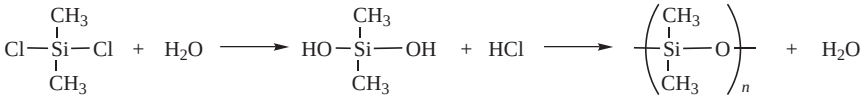


Figure 6. Formation of silicone prepolymer.

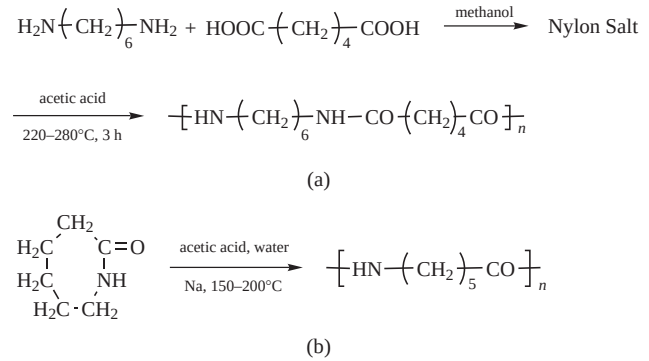


Figure 7. Synthesis of nylons: (a) nylon 66, (b) nylon 6.

direction is very high. The number and distribution of amide groups play an important role in determining the properties of nylons. Generally, the T_g and tensile properties increase with increasing the number of amide groups. For example, nylon 66 is stronger than nylon 610 and nylon 6 is stronger than nylon 11 (11a).

Nylon plastics are stiff and tough, and have a high abrasion resistance; but they are sensitive to water. Water adsorption reduces their strength and lowers their T_g. It is the amorphous region of polyamide chains that is sensitive to the attack of water. The greater the degree of crystallinity, the less the water adsorption. Nylons are used as surgical sutures, components of dialysis devices, hypodermic syringes, and intracardiac catheters.

Polyesters

Polyethylene Terephthalate. Polyethylene Terephthalate (PET) is known as Dacron. Commercially, it is manufactured by ester interchange polymerization between dimethyl terephthalate and excess glycol (1:1.7). The reaction has two stages (5d,12) (Fig. 8). In the first stage, methanol is displaced from dimethyl terephthalate by glycol. In the second stage, the excess glycol is driven off under vacuum. The polymer is semicrystalline with a T_m value of 265 °C and a T_g value of 80–120 °C depending on crystallinity. The PET fibers made from melt spinning have high strength and good crease resistance, so they are used as nondegradeable surgical sutures. Like PTFE, PET is also hydrophobic and hydrolysis resistant. The knitted or woven PET tubes are widely used for large diameter (12–30 mm) and medium diameter (5–11 mm) vascular grafts. However, Dacron graft devices are not fully satisfactory. Thrombosis is a major problem. Another drawback of Dacron grafts is the need for preclotting the grafts with autologous blood before implantation to prevent bleeding from their micropores (16a).

Polycarbonates. Bisphenol A polycarbonate is the only commercially significant polycarbonate product, so this material is often referred to as polycarbonate. It is prepared either by the reaction of phosgene with bisphenol A, or by ester interchange of a diphenyl carbonate with bisphenol A (Fig. 9). Polycarbonate is a clear, tough material. It has excellent mechanical and thermal properties (T_g 150 °C). The high transparency and impact strength make polycarbonate useful as lenses for eyeglasses and safety glasses, and housings for oxygenators and heart–lung bypass machines (2a). Requirements for medical grade polycarbonates are given in ASTM F 997.

Polyacetal

Polyacetal, also called poly(oxymethylene) is known as Delrin (DuPont). It is prepared from formaldehyde in an inert hydrocarbon solvent along with an initiator (ring-scission polymerization) (5d). The polymer has a high melting temperature (184 °C) and low glass transition temperature (−82 °C) (5b). It is lubricious, strong, and has good dimensional stability, resistance to creep and fatigue, high abrasion, and chemical resistance. A cementless polyacetal isoelastic femoral stem was introduced in the early 1970s to solve two important problems in total hip replacement (THR): stress shielding and cement disease. The modulus of elasticity of polyacetal (∼ 5–13 GPa) is close to that of bone, thus providing the condition of isoelasticity. However, the prostheses were unsuccessful in clinical practice due to high rate of loosening (20). The successful use of polyacetal in medical devices includes use in the valve disk in Penn State circulatory-assitant devices, which is one of only two approved for heart replacement by the U.S. Food and Drus admistration (FDA) (16d). Specifications for polyacetal are given in ASTM F 1855.

Hydrogels

Hydrogels are 3D networks of hydroplilic polymers held together by chemical or physical crosslinks. Typical methods to prepare hydrogels include irradiation, chemical reactions, and physical association. Hydrogels have inherently weak mechanical properties, so hydrophobic

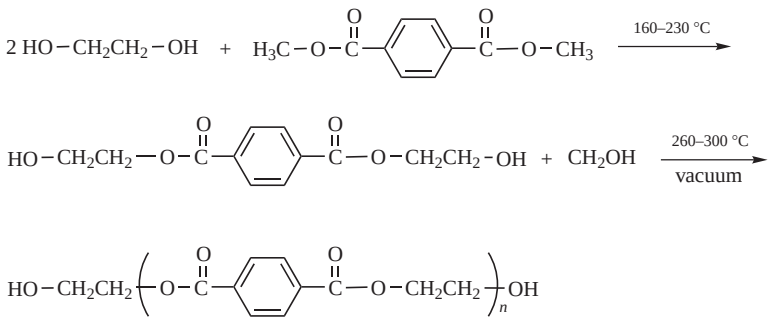


Figure 8. Synthesis of poly(ethylene terephthalate).

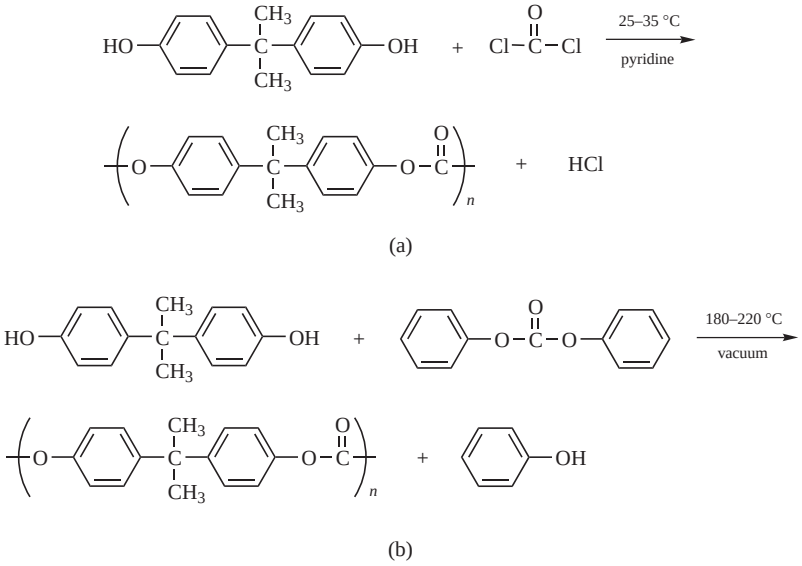


Figure 9. Synthesis of polycarbonate through (a) phosgene, (b) ester interchange.

constituents are often incorporated to improve the mechanical strength. The low interfacial tension with surrounding biological fluids makes hydrogels desirable for nonfouling surfaces (10c). Hydrogels tend to calcify under physiological conditions, limiting the use of hydrogels as implantable biomaterials (17c).

Poly(hydroxyethyl methacrylate) (PHEMA) is the most frequently used hydrogel. It is a rigid acrylic polymer when dry, but it takes up ~40% water when wet and changes into an elastic gel. PHEMA hydrogel is transparent, and thus is used for soft contact lens (17c). Other more hydrophilic hydrogel monomers, such as methacrylic acid and 1-vinyl-2-pyrrolidone, are often copolymerized with hydroxyethyl methacrylate to improve the oxygen permeability coefficient and water adsorption of PHEMA hydrogel. This hydrogel was also the first successfully used for wound dressings under the trade name Hydron (17c).

Poly(vinyl alcohol) is a water-soluble polymer. Solutions of PVA are used as ophthalmic lubricants and viscosity-increasing agents. The solution thickens the natural film of tears in eyes. Poly(vinyl alcohol) can crystallize even in its highly hydrolyzed state, and thus it has a relatively high tensile strength for a hydrogel. The PVA hydrogel is a candidate material for artificial articular cartilage in reconstructive joint surgery, and has been used for releasing bovine serum albumin. Physically cross-linked PVA hydrogels prepared by free-thaw processes have been investigated for protein-releasing matrix (17c).

Other synthetic hydrogels of biomedical interest include polyacrylamides, poly(*N*-vinyl-2-pyrrolidone), poly(methacrylic acid), and poly(ethylene oxide).

Biodegradable Synthetic Polymers

The interest in biodegradable polymers has dramatically increased in recent years, because these biomaterials do not permanently leave residuals in the implantation site, do not elicit permanent foreign-body reactions, and avoid second surgeries in the case of temporary implants like fracture fixation devices (10c,11b,17d). The major biomedical applications of biodegradable polymers include: temporary scaffolding and barrier, drug delivery matrix, tissue engineering scaffolds, and adhesives. Biodegradable polymers are generally hydrophilic. They degrade either by simple hydrolysis without enzyme catalysis or by enzymatic mechanisms. The degradation products should be inert or a natural metabolite of the body (11b).

The most commercially successful biodegradable polymers are polyglycolide (PGA), polylactide (PLA) and their copolymer poly(glycolide-*co*-lactide) (PLGA). The polymers PGA, PLA, and PLGA are a group of poly α-hydroxy acids belonging to absorbable polyesters. They degrade by bulk hydrolysis of their ester bonds. The hydrolysis byproduct of PLA is lactic acid, which is a normal byproduct of anaerobic metabolism in the human body (17e). The degradation of PGA involves both hydrolytic scission and enzymatic degradation, and the product is glycolic acid that can be eliminated by the metabolic pathway as carbon dioxide and water (1b).

Polylactide is prepared in the solid state through ring-opening polymerization. It has four stereoisomers: D-PLA, L-PLA, DL-PLA, and *meso*-PLA. The most frequently used forms in biomedical practice are L-PLA, and DL-PLA (2c). The L-PLA is a semicrystalline polymer with a T_m of 159–178 °C and T_g of 54–59 °C (17e). Compared with other biodegradable polymers, L-PLA exhibits high strength and modulus, and a very slow biodegradation rate. It is suitable for light load-bearing applications (e.g., orthopedic fixation devices, vascular grafts, and surgical meshes). Self-reinforced L-PLA bolts, screws, pins and anchors have been used for bone fracture fixation (1b). The DL-PLA is an amorphous polymer with a T_g of 51 °C. It degrades very fast, and thus usually is used for drug delivery (11b). The L-PLA polymer and its *in vitro* degradation tests are specified in ASTM F 1925 and F 1635, respectively.

Polyglycolide can be synthesized either by direct polycondensation of glycolic acid or by ring-opening polymerization of the cyclic dimers of glycolic acid. Due to tight molecular packing, PGA has a high melting point (225–230 °C). Its T_g ranges from 35 to 40 °C. The PGA degrades

faster than LPLA because it is more hydrophilic (11b) and PGA can be melt spun into fibers and fabricated into sutures, meshes, and surgical products. Dexon (American Cyanamid) is a trade name of polyglycolide products. It has been successfully used for wound closure and sutures (10c,11b). The Properties and degradation rate of PLGA can be controlled by varying the ratio of monomers. Vicryl from Ethicon is a poly(glycolide-L-lactide) random copolymer with 90:10 ratio of glycolide to lactide. It completely degrades *in vivo* after 90 days, and has been successfully used as surgical meshes and sutures, and for wound closure and drug delivery (10c,21b). Several other biodegradable polymers have been developed and investigated for drug delivery, tissue engineering, and medical devices [e.g., polyorthoesters, poly(ϵ-caprolactone), polydioxanone, and polyanhydride] (11b).

The concern with PLA and PGA is that their degradation products (lactic acid and glycolic acid) may significantly lower the local pH in a closed and less body-fluid-buffered region, leading to irritation at the site of polymer implant (11b).

Smart Polymers

Smart polymers are "polymers that respond with large property changes to small physical or chemical stimuli" (22). The most common stimuli are pH values and temperatures. Some polymers [e.g., poly(hydroxypropyl acrylate), poly(*N*-isopropylacrylamide), and poly(ethylene oxide/propylene oxide) copolymers] exhibit thermally induced precipitation and have a lower critical solution temperature (LCST). The polymers are soluble in water below LCST, but precipitate sharply as temperature is raised above LCST. Pluronic F127 (BASF, NJ) is a polyethylene oxide–polypropylene oxide–polyethylene oxide triblock copolymer with a LCST around the physiological temperature. It is used for controlled release of drugs including proteins and liposomes (23).

In general, pH-responsive hydrogels can be prepared from polymers with ionizable groups (e.g., carboxyl, sulfonic, amino, and phosphate groups). The pH change influences the ionization degree of the polymer to govern its solubility in water. Lysozyme (a cationic protein) immobilized within a hydrogel with phosphate groups is released at pH 7.4 (enteric conditions), but is not released at pH 1.4 (gastric conditions), ensuring the drug is delivered to the small intestines, and is not released in the stomach (17f).

Besides drug delivery, smart polymers can be used for sensors, chemical valves, mechanochemical actuators, specialized separation systems, and artificial muscles.

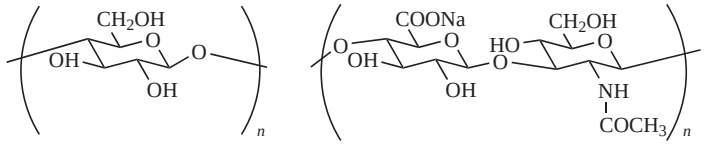
Biopolymers

Biopolymers are polymers of natural origin that can be obtained from animals, plants, and microorganisms. They are produced during the growth cycles of all organisms by enzyme catalyzed stepwise polymerization rather than a chain polymerization (5a). Biopolymers most frequently used for medical applications are polysaccharides and proteins.

Cellulose is a polysaccharide of plant origin. It is the primary structural component of plant cell walls. The molecular chain of cellulose is linear, consisting of D-glucose residues linked by β(1-4) glycoside bonds (Fig. 10a). The strongly hydrogen-bonded structure make cellulose highly crystalline and exceptional in strength, but insoluble in water and most organic solvents, and infusible. Ether and ester derivatives of cellulose, such as cellulose acetate and hydroxyethyl cellulose, have been developed to improve its processability (24). Cellophane (regenerated cellulose) semipermeable membrane was first used for hemodialysis to remove blood waste in the 1960s, and it is still in use today, but mostly in hollow fiber forms. Cellulose acetate (di- and triacetate) membranes and hollow fibers also find in hemodialysis. Hydroxypropylmethylcellulose is the most widely used hydrophilic drug delivery matrix (17f). The matrix tablets can be formed by compression, slugging, or wet granulation. Cellulose and its derivatives are also used as wound dressings.

Hyaluronan (HA) is a natural mucopolysaccharide, present in connective tissues of all vertebrates, which consists of repeating disaccharides of *N*-acetylglucosamine and glucuronic acid (Fig. 10b). Besides inherent biocompatibility, hyaluronan has some other unique properties: viscoelasticity, hydrophilicity, lubricity, and biological activity (regulator of cellular activity) (25). Unlike cellulose, hyaluronan is soluble in water forming a viscous solution, and it is biodegradable. Many derivatives have been developed to improve its residence time in water. Hylan (crosslinked hyaluronan, Biomatrix) and Hyaff (hyaluronan with carboxyl groups esterified by alcohol, Fidia Advanced Biopolymer) are two groups of widely used and commercialized hyaluronan derivatives. Native hyaluronan and Hylan are used for viscosupplementation to treat arthritis, viscoseparation to prevent adhesion and facilitate wound healing after surgeries, and viscoaugmentation to correct scars and facial wrinkles. Hyaff is widely used for wound dressings as sheets, meshes, or nonwoven fleeces. Hyaluronan and its derivatives have also been investigated for lubricious coatings of biomedical devices, control released drug delivery and tissue engineering scaffolds (26).

Collagens are a family of structural fiber proteins present in all animals (27). They are the most abundant proteins in



(a) (b)

Figure 10. Structure of polysaccharides: (a) cellulose, (b) hyaluronan.

vertebrates, widely distributed in pliant connective tissues and tensile structures (e.g., tendon and ligament) as scaffolds to keep their shape, maintain integrity, and support tensile stress. A collagen molecule is composed of three polypeptide chains wound into a triple helix, stabilized by interchain hydrogen bonds. They are subject to degradation by lysozomal enzymes and collagenase. Glutaraldehyde and carbodiimides can be used to crosslink collagens to improve their mechanical properties (10c). Collagens are probably the most common proteins used as biomaterials. They are used in artificial skins, in soft tissue, and plastic surgery to fill up tissue defects, in surgical sutures, vascular grafts, and corneal replacements (11c).

POLYMERIC BIOMATERIALS EVALUATION

Bulk Characterization

Initial characterizations give information about the composition, structure, and physical and mechanical properties of the investigated polymers to examine if the properties match the specific application requirements, and the reproducibility of batch-to-batch properties. Infrared (IR) spectroscopy and nuclear magnetic resonance (NMR) are often used to analyze the chemical composition and structures of polymers. Crystalline and multiphase structures are usually determined using wide- (WAXS) or small-angle (SAXS) X-ray scattering, or transmission electron microscopy (TEM). Differential scanning calorimetry (DSC), and dynamic mechanical analysis (DMA) can provide thermal transition information of polymers (e.g., T_m and T_g). Furthermore, DMA can provide significant insight into the viscoelastic nature of the polymer. The mechanical properties are determined using the standard ASTM methods. A second level of characterizations needs to be made on the candidate materials to investigate if sterilization and physiological environments' exposure significantly change their properties.

Surface Characterization

Biomaterials contact the body through their surface, and thus the surface chemistry and topography determine the host's response to the materials. The surface chemistry of polymers can be characterized with X-ray photoelectron spectroscopy (XPS), attenuated total reflectance Fourier transform infrared (ATR–FTIR) and secondary ion mass spectroscopy (SIMS). Scanning electron microscopy (SEM) and atomic force microscopy (AFM) are typical techniques to determine the surface topography. A contact angle goniometer is used to measure the wettability (i.e., hydrophilicity) of surfaces.

Biocompatibility Assessment

Biocompatibility tests generally include two levels. The first level tests are biosafety testing, while the second level involves biofunctional testing (28). In biosafety tests, the materials or their extracts are tested to see if they are toxic to cultured cells, cause hemolysis or allergic responses, induce heritable genetic alterations, or tissue necrosis after animal implantation. Those materials passing the first level tests need to be further inspected with the second level tests. This level of testing focuses on the specific functions of a medical device, in which the responses of all the cell and tissue types contacting the device are investigated with both *in vitro* and *in vivo* methods. The functionality of the medical device is also tested during this phase of testing, and changes in material can have significant effects on functionality just as changes in medical device design or function can have significant effects on the biocompatibility of the material.

BIBLIOGRAPHY

Cited References

1. (a) Ikada Y. Polymeric biomaterials in medical systems. (b) Mauliagrawal C. Biodegradable polymers for orthopaedic applications. In: Reis RL, Cohn D, editors. Polymer Based Systems on Tissue Engineering, Replacements and Regeneration. Dordrecht: Kluwer Academic Publishers; 2002.
2. (a) Cooper SL, Visser SA, Hergenrother RW, Lamba NMK. Polymers. (b) Hallab NJ, Jacobs JJ, Katz JL. Orthopedic applications. (c) Kohn J, Abramson S, Langer R. Bioresorbable and bioerodible materials. In: Ratner BD, Hoffman AS, Schoen FJ, Lemons JE, editors. Biomaterials Science: An Introduction to Materials in Medicine. 2nd ed., San Diego: Academic Press; 2004.
3. Ikada Y. Biological Materials. In: Barbucci R, editor. Integrated Biomaterials Science. New York: Kluwer Academic/Plenum Publishers; 2002.
4. Cheng HN, Gross RA. Polymer Biocatalysis and Biomaterials. In: Cheng HN, Gross RA, editors. Polymer Biocatalysis and Biomaterials. Washington, (DC): American Chemical Society; 2005.
5. Rodriguez F, Cohen C, Ober CK, Archer LA. Principles of Polymer Systems, 5th ed., New York: Taylor & Francis; 2003. Chapt 4(a); Chapt 3(b); Chapt 15(c); Chapt 16(d).
6. Young RJ. Introduction to Polymers. London: Chapman and Hall; 1981. Chapt 2.
7. Matyjaszewski K. Comparison and classification of controlled/living radical polymerizations. In: Matyjaszewski K, editor. Controlled/Living Radical Polymerization: Progress in ATRP, NMP, and RAFT. Washington, (DC): American Chemical Society; 2000.
8. Chanda M. Advanced Polymer Chemistry: a Problem Solving Guide. New York: Marcel Dekker; 2000; Chapt 1(a); Chapt 2(b).
9. Askeland DR, Phule PP. The Science and Engineering of Materials. 4th ed., Boston: PWS Publishing Company; 2003; Chapt 15.
10. Bhat SV. Biomaterials. Boston: Kluwer Academic Publishers; 2002; Chapt 12(a); Chapt 8(b); Chapt 5(c); Chapt 9(d); Chapt 6(e).
11. (a) Lee HB, Khang G, Lee JH.. Polymeric biomaterials. (c) Chu CC. Biodegradable polymeric biomaterials: an updated overview. (d) Li ST. Biologic biomaterials: tissue-derived biomaterials (collagen). In: Park JB, Bronzino JD, editors. Biomaterials: Principles and Applications. Boca Raton (FL): CRC Press; 2003.
12. Alger MSM. Polymer Science Dictionary. London: Elsevier Science Publishers; 1989.
13. Birnkraut HW. Synthesis of UHMW-PE. In: Willert H-G, Buchhorn GH, Eyerer P, editors. Ultra-High Molecular Weight Polyethylene as Biomaterial in Orthopedic Surgery. Toronto: Hogrefe & Huber Publishers; 1991.

14. Kurtz SM, Muratoglu OK, Evans M, Edidin AA. Advances in the processing, sterilization of ultra-high molecular weight polyethylene for total joint arthroplasty. Biomaterials 1999;20:1659–1688.
15. Teoh SH, Tang ZG, Hastings GW. Thermoplastic polymers in biomedical applications: Structures, properties and and processing. In: Black J, Hastings G, editors. Handbook of Biomaterial Properties. London: Chapman & Hall; 1998.
16. (a) Ramshaw JAM, Werkmeister JA, Edwards GA. Tissue-polymer composite vascular prostheses. (b) Kowligi RR, Edwin TJ, Banas C, Calcote RW. Vascular grafts: materials, methods, and clinical applications. (c) Planell JA, Vila MM, Gil FJ, Driessens FCM. Acrylic bone cements. (d) Felder G III, Donachy JH, Sr. Fabrication techniques and polymer considerations for the blood contacting components of the Penn State circulatory–assist devices. In: Wise DL, Trantolo DJ, Altobelli DE, Yaszemski MJ, Gresser JD, Schwartz ER, editors. Encyclopedic Handbook of Biomaterials and Bioengineering, Part B: Applications Vol. 2. New York: Marcel Dekker; 1995.
17. (a) Tomita N, Fujita H, Nagata K. Polymers for artificial joints. (b) El-Zaim HS, Heggers JP. Silicones for pharmaceutical and biomedical applications. (c) Kishida A, Ikada Y. Hydrogels for biomedical and pharmaceutical applications. (d) Rokkanen PU. Bioabsorbable polymers for medical applications with an emphasis on orthopedic surgery. (e) Domb AJ, Kumar N, Sheskin T, Bentolila A, Slager J, Teomim D. Biodegradable polymers as drug carrier systems. (f) Miyata T, Uragami T. Biological stimulus-responsive hydrogels. (g) Dumitriu S. Polysaccharides as biomaterials. In: Dumitriu S, editor. Polymeric Biomaterials, 2nd ed. New York: Marcel Dekker; 2002.
18. Lamba NMK, Woodhouse KA, Cooper SL. Polyurethanes in Biomedical Applications, Boca Raton (FL): CRC Press; 1998. Ch. 2.
19. (a) Szycher M, Reed AM. Biodurable polyurethane elastomers. (b) Tiffany JS, Petraitis DJ. Silicone biomaterials. In: Wise DL, Trantolo DJ, Altobelli DE, Yaszemski MJ, Gresser JD, Schwartz ER, editors. Encyclopedic Handbook of Biomaterials and Bioengineering, Part A: Materials (Vol. 2). New York: Marcel Dekker; 1995.
20. (a) Pappas MA, Schmidt CC, Shanbhag AS, Whiteside TA, Rubash HE, Herndon JH. Biological response to particulate debris from nonmetallic orthopedic implants. (b) Gresser JD, Trantolo DJ, Lyons CH, Nagaoka H, Shuster L, Swift RM, Wise DL. In vitro and in vivo release of naltrexone from two types of poly(lactide-co-glycolide) matrices. In: Wise DL, Trantolo DJ, Altobelli DE, Yaszemski MJ, Gresser JD, editors. Human Biomaterials Applications. Totowa (NJ): Humana Press; 1996.
21. Minovic A, Milosev I, Pisot V, Cor A, Antolic V. Isolation of polyacetal wear particles from periprosthetic tissue of isoelastic femoral stems. J Bone Joint Surg 2001; 83B:1182–1190.
22. Hoffman AS, Stayton PS, Bulmus V, Chen G. Really smart bioconjugates of smart polymers and receptor proteins. J Biomed Mater Res 2000;52:577–586.
23. Chandaroy P, Sen A, Hui SW. Temperature-controlled release from liposomes encapsulating Pluronic F127. J Controlled Release 2001;76:27–37.
24. Klemm D, Philipp B, Heinze T, Heinze U, Wagenknecht W. Comprehensive Cellulose Chemistry Vol. 2: Functionalization of Cellulose. Weinheim, Germany: Wiley-VCH; 1998.
25. Laurent TC. The Chemistry, Biology and Medical Applications of Hyaluronan and its Derivative. London: Portland Press; 1998.
26. Kennedy JF, Philips GO, Williams PA. Hyaluronan 2000. Cambridge (England): Woodhead Publishing Limited; 2002.
27. Wainwright SA, Biggs WD, Currey JD, Gosline JM. Mechanical Design in Organism. Princeton (NJ): Princeton University Press; 1982. Chapt 3.
28. Zhang M. Biocompatibility of materials. In: Shi D, editor. Biomaterials and Tissue Engineering, Heidelberg: Springers; 2003. p 83–137.

See also BONE CEMENT, ACRYLIC; CONTACT LENSES; LENSES, INTRAOCULAR; POLYMERIC MATERIALS.

BIOMATERIALS, SURFACE PROPERTIES OF

SALLY L. MCARTHUR
ALEXANDER G. SHARD
University of Sheffield

INTRODUCTION

In the broadest of definitions, biomaterials are nonliving materials that come into contact with biological systems. The point of contact between the two different phases is at the interface, or surface, of the material. It is quite common for the surface of a material to have properties that are not trivially related to the bulk of the material. These differences can arise because of a number of processes, such as surface segregation, surface reactions, contamination, scratching, and phase separation. It should therefore be recognized that the interactions between a biomaterial and the biological medium and in turn, the physical and chemical activity or stability of a medical device, can depend critically upon the properties of the surface.

In general, materials selection for biomedical devices and applications is based on a combination of physical properties, manufacturability, and availability. In many cases, materials are chosen because they have been used previously in medical devices and as such, detailed records of *in vivo* behavior and performance already exist. Due to the costs and time involved with the testing of new materials to meet regulatory standards for safety and efficacy, relatively few materials are currently used in the manufacture of biomedical devices. The most common of these are titanium-based alloys, 316L stainless steels, ultrahigh molecular weight polyethylene (UHMWPE), expanded poly(tetrafluoroethylene) (e-PFTE), poly(ethylene terephthalate) (PET), poly(hydroxyethyl methacrylate) (pHEMA), polyglycolic and lactic acids (PGA and PLA), polystyrene, polyurethanes, hydroxyapatite, alumina, and zirconia.

Of course, mechanical properties are only one of a number of materials characteristics that may be required for biomedical applications. Each biomedical application may desire a range of properties that are directly influenced by the nature of the surface. Specific characteristics and modifications made to biomaterial surfaces include:

1. Orthopedic Devices

Improved wear resistance and frictional properties for joints and bearing surfaces via cross-linking of UHMWPE and the introduction of carbide, nitride and crystalline structures on metallic components.

Bone conductive coatings for improved osseointegration via implantation of specific chemical species (e.g., Ca, P) into metals and deposition of hydroxyapatite coatings

2. Cardiovascular Devices

Improved hemocompatibility via diamond-like carbon (DLC) coatings (e.g., mechanical heart valve leaflets) and via the immobilization of biomolecules to promote epithelialization

Short- and long-term drug delivery via degradable polymeric coatings (e.g., stents).

Polymeric barrier coatings to prevent transmission of electrical signals and improve corrosion resistance (e.g., pacemaker cases and leads).

3. Diagnostics, Sensors and *in vitro* Applications,

Reflective coatings for optical sensors.

Nonfouling coatings to prevent protein and cell attachment and reduce background signal in biological assays and sensors.

Oriented biomolecule immobilization for DNA, protein, and antibody arrays.

Topographical and chemical patterning of microfluidic devices and sensors for the control of fluid flow and chemical mixing.

4. Tissue Engineering

Improved cell proliferation and growth in culture via oxidation of polystyrene to produce a hydrophilic substrate (tissue culture polystyrene, TCPS).

Immobilization of biological ligands for controlled cell adhesion (e.g., RGD and other cell receptor binding domains).

In this article, we intend to provide a broad overview of the basic properties of surfaces, their interactions with biological systems, and how surfaces can be changed to suit particular biomaterial applications. Of particular importance is the requirement for surface characterization. As stated earlier, differences between surface and bulk properties can arise via a number of different processes. However they arise, it is important to ensure that the surface properties of the material are verified before ascribing any biological effect to the material. To complete this article, we provide an outline of the most commonly employed surface characterization techniques and include references to more detailed texts to aid the interested reader.

PROPERTIES OF SURFACES

One of the most important properties of a surface or interface is that it exhibits free energy. This means that if the surface was extended in some way so that it had a larger area then work would have to be done. If this was not the case, then for fluid interfaces at least, the surface could grow without limit, eventually resulting in a homogenous mixture. The existence of surface energy leads to a tendency for surfaces to contract resulting in a higher pressure on the inside of curved surfaces. Measuring the interfacial energy between liquids and air is relatively trivial, as the surface may be extended without producing a bulk strain. Thus the energy required to extend an area of surface or, more usually, the force of contraction normal to a length of surface can be directly obtained. Liquid surface tensions scale with the strength of intermolecular interactions in the bulk of the liquid, so hydrocarbons typically have surface energies of $\sim$25 $mN \cdot m^{-1}$, water has 72 $mN \cdot m^{-1}$ due to hydrogen bonding and the metallic bonding in mercury results in a surface energy of over 470 $mN \cdot m^{-1}$ (1).

In contrast, the surface energy of solids cannot be obtained directly. There are, however, a variety of methods of estimating it from a series of contact angle experiments and it is found that the surface energies between solids and air are very similar to those of analogous liquids. However, for most biomaterial applications it is the solid–water interfacial energy that is important. One should note that "low energy" hydrophobic surfaces typically have interfacial energies with both air and water of $\sim$30 $mN \cdot m^{-1}$. Hydrophilic surfaces, such as clean glass or aluminium, can have high surface energies in air, higher than 80 $mN \cdot m^{-1}$, but have negligibly small interfacial energies with water. The difference between the air and water interfacial energies for glass and aluminium is greater than the surface energy of water, and hence water does not form drops, but completely wets these materials. Protein adsorption, described in detail later, can be thought of as being driven by the minimization of surface energy. A comparison of the surface energies for hydrophilic and hydrophobic materials gives an appreciation of the strength and importance of the hydrophobic interaction, described later, during this process.

Other properties of surfaces that are important are chemistry, mechanical properties, and topology. In the context of a biomaterial, the chemistry of a surface will determine the initial interactions with proteins through ionic, hydrogen bonding, and hydrophobic interactions as well as the promotion of specific interactions by the presence of surface bound ligands. It is important to realize that the surface chemistry of a material may bear little or no resemblance to the bulk chemistry. In many cases, this is due to the presence of thin layers of contaminants that naturally accumulate on the surfaces of all materials. The deliberate alteration of biomaterial surface chemistry is carried out to enhance or inhibit certain properties, usually the alteration of protein adsorption and cell attachment. Whether the surface chemistry is a result of contamination or modification, it is important to specifically characterize the surface to ensure that correlations between biomaterial chemistry and performance are correctly obtained. The mechanical properties of a biomaterial surface are also of some importance, particularly for cellular attachment. It is generally found that cells attach more strongly to rigid substrates and will migrate from soft-to-hard materials. Note that the mechanical properties of some materials, in

particular polymers, may be somewhat different at the surface compared to the bulk. It has been observed, for example, that the glass transition temperature of polymers is reduced close to an interface. The topology of a surface is also important as it defines the surface area of the interface, and has been shown to influence cell behavior (2).

The Adsorption of Proteins at Surfaces

One of the most important events that occur in biomaterial applications is the sequestration of proteins from solution to the surface of the material. Proteins are polyamino acids in which, for each protein, there is a predetermined and specific sequence of amino acids. This sequence is termed the primary structure of the protein. The secondary structure consists of a variety of common folding motifs, such as α-helices and β-sheets. The tertiary structure of the protein comprises the folding and packing of the secondary structure into a particular three-dimensional (3D) shape. For most proteins, the tertiary structure creates unique, and often rather small sites of activity that allow the protein to function (e.g., cell-binding domains). In contrast, synthetic macromolecules form random coils because they lack the well-defined structure that allows the strong bonding that occurs between different parts of the protein chain.

When one considers protein adsorption at interfaces, it is common to draw analogies to the adsorption of synthetic macromolecules. While these comparisons are extremely useful, it is important to remember that proteins are capable of site specific and highly selective binding, whereas synthetic macromolecules in general are not. Examples of such selective binding include the much utilized affinity of avidin for biotin and the binding of antigens to antibodies. Protein adsorption occurs primarily due to a number of intermolecular forces. These include ionic and hydrogen bonding and the hydrophobic interaction (3). Although the ionic interaction is rather strong in solid materials, in aqueous media it is diminished due to strong ion–dipole interactions with water, the high dielectric constant of water and the presence of other solvated ions that cause a decrease in the effective range of ionic interactions. Nevertheless, ionic interactions are important at short ranges and can have a strong effect on the rate of adsorption of proteins at surfaces. It is commonly observed, for example, that a protein adsorbs most rapidly to an uncharged surface when it is at its isoelectric point, that is, when it is itself uncharged. The presence of a charged interface can decrease or increase the rate of adsorption depending on whether the protein has a like or an unlike total charge. Furthermore, if the protein has a dipole moment, then it may be possible to influence the orientation of the protein upon adsorption at a charged surface.

Hydrogen bonding is a particularly strong example of a dipole–dipole interaction. A hydrogen atom bound to an electronegative element such as oxygen or nitrogen forms a strong association with a lone pair of electrons on another electronegative atom, which may be part of another molecule. There is no great driving force for the formation of hydrogen bonds in the presence of water, since water very effectively makes such bonds. Without the generation of highly specific geometries of complementary hydrogen donors and acceptors, hydrogen bonding is almost certainly not a major driving force for adsorption of proteins at interfaces.

The "hydrophobic interaction" is something of a misnomer, since the driving force is in fact the formation of hydrogen bonds in water and not the attraction between two hydrophobic species. Water cannot form hydrogen bonds with regions of predominantly hydrocarbon species, whether these are part of a protein or on a surface. The result is that at such an interface water is in a state of higher free energy than if the interface was not present. Hydrocarbons thus tend to aggregate together to minimize the area of contact between themselves and water and lower the free energy of the system as a whole. These interactions are critical to the folding of proteins, with the interior of the protein generally consisting of hydrophobic amino acids and the exterior of hydrophilic amino acids. It is undoubtedly also an important interaction in the adsorption of proteins at interfaces. While the exterior surface of most globular proteins contain few hydrophobic sites, if the protein can unfold upon the surface (denature) then many more such sites become available.

When a surface is exposed to a solution of a single protein it is generally found that adsorption occurs rapidly and in many cases is diffusion limited. It is usual for adsorption to reach a maximum at a single layer with close contact between adsorbed proteins. Following adsorption, the rate of desorption from the surface is extremely slow. Proteins cannot commonly be removed from surfaces simply by changing the protein solution for pure solvent. However, if other proteins are present there may be exchange between adsorbed and solvated proteins. This includes self-exchange, as has been demonstrated by the exchange of unlabelled proteins with their radiolabeled analogues (4). Different proteins can have different affinities for surfaces, so that one protein may adsorb initially because it is in a high solution concentration, but at later times be displaced by other proteins that have higher affinity, but are in low concentrations. This effect is named after Leo Vroman and the classic example is the adsorption of proteins from serum that occurs in the order albumin, fibrinogen, and high molecular weight kininogen. It is also thought that immunoglobulin G adsorbs transiently between albumin and fibrinogen (5). This exchange can take place in a matter of seconds in pure serum, but may take minutes or hours in diluted serum. It is also noted that the amount of protein that can be exchanged in this manner diminishes the longer the protein is in contact with the surface. This indicates that the initial state of adsorption is metastable and that some activation energy barrier needs to be overcome for an adsorbed protein to reach a free energy minimum. It is possible that this energy barrier relates to the unfolding of tertiary or secondary structure and represents a denaturation of the protein.

Although the precise details of protein adsorption are unclear, it is generally agreed that the stability of adsorbed protein layers derives from the large number of contact points possible between a single protein molecule and a surface. Although each individual contact may be weak and temporarily displaced by smaller molecules the probability of breaking enough bonds for the protein to actually desorb is extremely small. The stability of an adsorbed layer is

therefore related to both the strength of individual interactions with the surface and the number of interactions. One should expect on this basis that, neglecting the detailed interactions and protein conformation, a high molecular weight protein should displace a low molecular weight protein because it is able to form more bonds to the surface. It is instructive to note that this trend is at least partially followed in the Vroman effect, the exception being high molecular weight kininogen that has a slightly lower molecular weight than fibrinogen.

Cell Behavior at Surfaces

In comparison with protein adsorption, the adhesion of cells to a biomaterial surface is a rather slow process. In standard cell culture, the adsorption and equilibration of proteins at the surface will occur much more rapidly than cellular attachment. The behavior of cells at a surface is thought to be governed by the initial layer of protein. Cells with surfaces via interactions of their transmembrane proteins (e.g., integrins) with proteins in the extracellular matrix. One approach to encourage cell adhesion is to incorporate such specific sequences at the surface of the biomaterial. A variety of suitable peptide sequences have been reported. From fibronectin, the RGD sequence mentioned above and also REDV, which targets integrins found in endothelial cells, but not blood platelets. Laminin contains sequences such as YIGSR and SIKVAV, which may be employed to encourage nerve cell growth (6).

If cell attachment and growth is to be discouraged, then the biomaterial surface should adsorb as few proteins as possible or only adsorb proteins that are not implicated in cellular adhesion. In the first alternative, this is typically achieved by using a hydrogel-like polymer layer, such as grafted chains of polyethylene glycol. These highly hydrated films provide few sites for protein attachment and cell attachment is also strongly discouraged. It is commonly observed that cells attach poorly to hydrophobic surfaces; this may indicate that there is a selective adsorption of proteins that do not contain binding domains for cells. The modification of surfaces to promote and inhibit cell attachment is discussed later in this article.

Once a cell has formed attachment points at a surface it will strengthen these by accumulating integrin receptors in the vicinity of each site. These eventually form a focal adhesion that acts as a connection between the actin cytoskeleton of the cell and the surface. As these adhesive contacts are made the cell spreads upon the surface and will then enter the normal cell cycle. The formation of focal adhesions is critical to the survival of the cell, without sufficient spreading a cell will normally die. There are proteins that trigger signals from the focal adhesion to the cellular interior such as focal adhesion kinase, which may be implicated in this decision making process.

The movement of mammalian cells is achieved by crawling. This involves the myosin driven contraction of actin filaments in the cell to supply the mechanical power, the detachment of focal adhesions at the trailing edge of the cell and the formation of new adhesions at the leading edge (7). Cells will generally move in the direction in which they can make the largest number of focal adhesions. The surface of a biomaterial can thus be tailored to concentrate cells in particular locations.

SURFACE MODIFICATION

In many cases, surface characteristics can be modified by designing the chemical constituents of the materials, for example, surface segregating components in polymer blends to alter frictional properties; or induced during the manufacturing process, for example, the introduction of topography via die and mould design. However, it is not always possible or practical to use these approaches and secondary processing capable of inducing specific surface properties without detrimentally affecting the bulk characteristics is often required.

In broad terms, surface modification techniques can be divided into two categories: those that treat the existing surface and those that result in the addition of a surface coating. As shown in Table 1, there are a number of different surface modification techniques that are currently used in industry or applied to bioengineering research. In this section, we give a brief overview of a number of these techniques, discuss their advantages, and limitations and give some specific examples of their application.

Plasma Treatment and Polymerization

Plasma-based modifications have been applied, with varying degrees of success, to biomaterials and biomedical devices since the early 1960s. Also termed radio frequency glow discharge (rfgd), the process involves the volatilization of a liquid or gaseous "monomer" into an evacuated process chamber. An electric field at rf is applied across the vapor, ionizing a fraction of the molecules and generating electrons, ions, free radicals, photons, and molecules in both ground and excited states, within the gas plasma. When the resultant reactive species impinge on a surface within the plasma zone, they create reactive sites resulting in alteration of the surface chemistry and properties.

There are two classes of glow discharge plasma modification, treatment and Polymerization. Plasma treatment results in the introduction of chemical species or physical changes to the surface of the material. Plasma treatments are often used to etch polymeric, metallic and ceramic surfaces, remove contaminants, and improve adhesion and hydrophilicity (8). Chemical modifications resulting from plasma treatments can also be used as an activation step for graft polymerization. Plasma generated radicals can be used to initiate polymerization of monomers in the liquid or gas phase, resulting in surface -grafted polymer layers. Typically "monomers" used for plasma treatment include oxygen, argon, ammonia, air, and water.

Plasma Polymerization occurs when a plasma is struck in an organic vapor and results in the deposition of a polymeric film from the vapor phase. Excitation of the monomer results in reactive species impinging on a surface within the plasma zone creating reactive sites that are then used for the covalent attachment of other species and subsequent growth of a coating of controllable thickness (typically tens of nanometers). A wide range of monomers can be used to produce plasma polymer coatings suitable

Table 1. Methods and Applications of Surface Modification Commonly Used in Biomedical Devices

Method	Application	References
Plasma polymerization	Organic and inorganic coatings for use as barrier coatings (thermal and chemical). Improved abrasion resistance, electrical and optical properties. Control of chemical functionality, cell and protein adhesion	8,9
Plasma treatment	Introduction of chemical functionality, crosslinking of polymers for improved wear and frictional properties	9,10
Plasma immersion or source ion implantation (PIII)	Wear resistance and improved friction properties for metals ceramics and polymers. Improved biocompatibility	11–13
Radiation techniques [ultraviolet (UV), gamma and laser irradiation]	Polymer grafting, introduction of topographical features and chemical functionality	14–17
Ion implantation	Improved wear and friction properties. Implantation of specific elements can improve cellular integration on polymers and metals	18–20
IBAD	Enhanced cell and tissue compatibility, antimicrobial properties, friction, wear, and chemical stability.	20–22
Polymer grafting	Nonfouling and biomimetic surfaces. Control of hydrophilicity, introduction of chemical functionality. Chemical, thermal, and biologically responsive coatings	23–25
Biomolecule immobilization	Biomimetic surfaces, introduction of specific biological function and activity.	26–29

for biomaterials applications. Table 2 lists some of the most common monomers and their applications. In general, plasma polymers tend to be highly cross-linked and do not reproduce the chemistry of the monomer. In the last 10 years, there has been increasing interest in the production of plasma polymers with the functionality and specific characteristics of their parent monomer. This can range from simple systems for retaining more amine or acid functionality in coatings (30) to more complex cases such as optimizing the protein resistance of poly(ethylene oxide)-like plasma polymers (31) or the production of thermally responsive *N*-isopropylacrylamide (NIPPAm) surfaces (32).

A range of deposition parameters can be used to manipulate the characteristics of a plasma polymer and encourage coating properties that are commensurate with those of a traditionally synthesized polymer. Lower deposition powers, pulsing of the power supply, and copolymerization have all been used to modify the coating properties (30,33,34). The resulting materials have been shown to retain higher monomer functionality and in some cases specific physicochemical properties normally associated with multi-step polymer grafted surfaces (32).

Ion Implantation and Ion-Beam Assisted Deposition

As is the case with plasma techniques discussed previously, the key difference between these two ion-based surface modification techniques is that ion implantation is a surface treatment while ion-beam assisted deposition (IBAD), as the name suggests, results in a surface coating. In both cases ionized species are produced via an ion source and accelerated in an electric field to reach the surface with kiloelectronvolt energies. Parameters affecting the process include beam energy, dose, and current density as well as the nature of the ion species (20).

In polymers, ion impacts and interactions induce modifications of the macromolecular structure through gas evolution, formation of double bonds, chain scissions, and cross-linking over a thickness corresponding to the penetration depth of the ions. Factors such as chain scission and cross-linking obviously have diametrically opposing effects on the properties of the polymeric surface. Generally, manufacturers utilize ion-beam implantation to increase cross-link density, a factor that can improve wear properties at load bearing interfaces and create polymers with improved chemical resistance. In metals and ceramics, ion implantation can be used to induce the formation of new surface phases, surface disorder. The formation of hard-phase nitride, carbide, and oxide precipitates via ion implantation has been used to harden the surfaces of Ti and Ti alloy orthopedic implants, improving their wear resistance (35). The application of specific ion species such as nitrogen and calcium enables the generation of specific chemical

Table 2. Common Plasma Polymerization Monomers and the Coating Properties They Produce

Monomer	Coating Properties and Applications
Organosilanes (silanes and disiloxanes)	Thermal and chemical resistance Specific electrical and optical properties
Fluorine (e.g.) and hydrocarbon (octadiene) containing	Hydrophobic coatings Chemical barrier coatings, non cell adhesive.
Acid containing (acrylic acid)	Hydrophilic coatings
Amine containing (heptylamine, allylamine)	Acid and amine functionality used for polymer and biomolecule immobilization Controlled cell attachment and growth
Ethylene oxide containing (glymes, diethylene glycol vinyl ether)	Nonfouling coatings Controlled protein and cell adhesion

changes at the surface. The bone conductivity, corrosion and wear resistance of Ti alloys have all been shown to improve after calcium and phosphorous ion implantation (35). On polymeric materials, nitrogen ion implantation has been used to induce complex crosslinked surfaces with increased solvent and wear resistance (36), while the incorporation of silver (Ag) ions has been used to impart antimicrobial properties on indwelling catheters (21). Ion-beam assisted deposition, combines ion-beam implantation with physical vapor deposition (PVD), producing a low stress, uniform and adherent coating via interactions of the ions from the beam with the coating atoms (20). Ion-beam assisted deposition (IBAD) has been used to produce a variety of metallic and inorganic coatings on Co–Cr and Ti alloys, alumina, and UHMWPE (37). Titanium alloy coatings have been produced on Co–Cr components to improved cellular integration in orthopedic applications (38) and bone conductivity has been improved on a variety of metallic substrates with the deposition of adherent hydroxyapatite coatings. Commercially, IBAD is used to produce DLC coatings that are chemically inert, optically transparent, have a low friction coefficient, and are extremely hard. These DLC coatings are used to treat the bearing surfaces of orthopedic implants to improve wear and friction properties and reduce the incidence of wear debris (39). On polymeric substrates, IBAD is used to produce adhesive silver coatings for antimicrobial applications (40).

Plasma Immersion Ion Implantation

Plasma immersion ion implantation (PIII) has a critical advantage over the standard ion implantation methods discussed in the previous section: it is not a line of sight technique, a factor that enables the modification of complex shapes commonly found in biomedical applications. Unlike ion implantation, PIII samples are pulse-biased to a high negative potential relative to the chamber wall and surrounded by high density plasma. Ions generated in the plasma are accelerated across the sheath formed around the samples and are implanted into the surface. Gaseous plasmas can be induced using a variety of sources including rf and microwave, and combining these gas plasmas with a metallic plasma allows interface mixing that result in metallic coatings with low intrinsic stress, significantly reducing the risk of coating delamination (12). Plasma immersion ion implantation has been used to surface modify skeletal prosthetic implants with Ti alloy coatings (for cell recruitment) while maintaining the mechanical properties of the Co–Cr substrate (41) and to deposit carbon and Ti–N coatings on both metals and polymers for improved wear and scratch resistance (12).

Wet Chemical Techniques

There is a vast array of wet chemical techniques that can be used to modify the surfaces of biomaterials. In their simplest incarnation, wet chemical routes for surface modification can involve the immersion of a device in a chemical bath to adsorb polymer to the surface. The complexity of the modification increases incrementally through the grafting of polymers to form nonfouling and bioresponsive coatings toward the construction of biomimetic surfaces that attempt to imitate the outer surface of a cell and can contain a range of lipids, proteins, and sugars.

Grafted polymer layers can be used to manipulate both the physical and the chemical characteristics of a surface. Grafting can be achieved via a number of routes including covalent coupling, surface graft polymerization; surface segregation and interpenetration of a substrate. One of the most popular current applications is in the generation of nonfouling surfaces via the immobilization of water-soluble polymers like polyethylene oxide (PEO). While there is considerable debate on the efficacy of these coating, recent reviews on surface modification for nonfouling behavior have detailed the critical roles of polymer molecular weight, graft density, and residual charge on the performance of these types of grafted polymer layers (23,42).

An alternative approach to polymer grafting is the self-assembly of molecules to form monolayers. Self-assembled monolayers (SAMs) can be formed spontaneously via a range of specific molecule–surface interactions. Common systems include alkane thiol on gold or silver and chlorosilanes on hydroxyl-terminated surfaces. By tailoring the headgroup chemistry of the immobilized molecules, surface can be designed with a range of properties. Commercially, these systems are currently used as platforms for biosensors and bioarray technologies (43). The well-defined nature of these systems has resulted in their extensive application in research as model systems for protein and cell–surface interaction studies (42).

Increasing focus on the development of coatings capable of eliciting specific biological responses has seen a significant research focus on the incorporation of peptide sequences, particularly from the receptor-binding domains of adhesion proteins, in order to promote cell adhesion (6,27). In more general terms, there is significant interest in the immobilization of a range of biomolecules. Array and sensor technologies require antibody, protein, and DNA immobilization, while the immobilized proteoglycans such as heparin have been used to the modulate hemocompatibility of biomedical devices (44). As illustrated in Fig. 1, immobilization strategies for biomolecules can be as simple as nonspecific adsorption or as complex as molecular imprinting. The adsorption of biomolecules tends to result in coatings with limited functionality as the molecules are randomly oriented and tend to be desorbed from the surface over time. Covalent immobilization can eliminate problems associated with desorption, but there is often little control over conformation or orientation and thus activity of the biomolecule can be limited. The use of spacer polymer chains or amino acid sequences between the surface and the protein can reduce the denaturation of the molecule. Examples of this type of approach are the site specific modification of proteins with cysteine that enable immobilization of the proteins in specific orientations on gold (45). The success of strategies designed to present biological ligands can also be maximized if the immobilized molecule is coupled to a surface capable of preventing nonspecific adsorption. In general terms, biological response is influenced by the presentation, average density and the spatial distribution of the immobilized molecule

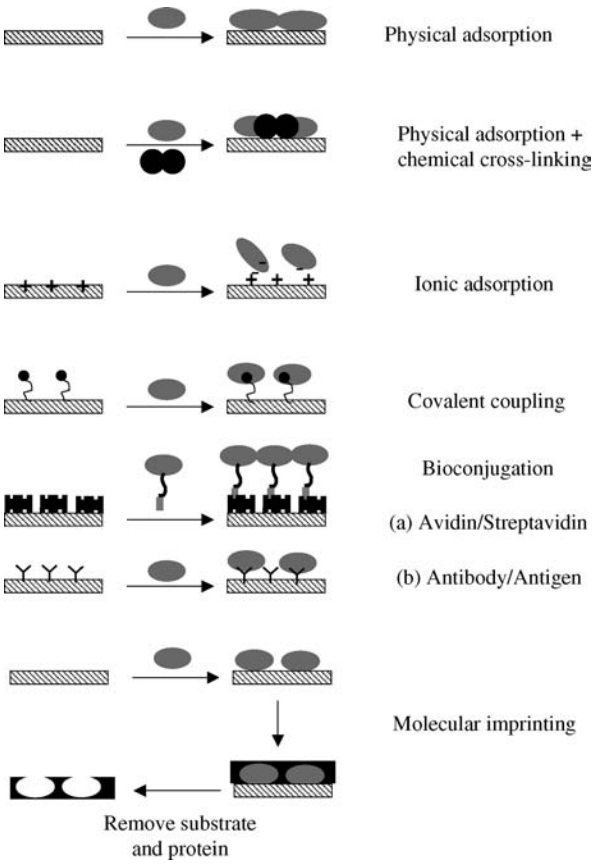


Figure 1. A range of immobilization strategies for biomolecules at interfaces.

(1). One of the most challenging applications for wet chemical surface modification lies in the development of surfaces that borrow from structures observed in Nature. Surfaces that mimic the structure of a cell wall are increasingly sought for use in biosensors and as model systems to further investigate cell, protein, and pharmaceutical interactions. These systems generally consist of a lipid bilayer that may containing a range of different lipids, transmembrane and membrane proteins, and in some cases oligosaccharides. Critically, these structures need to retain their fluidity; molecules need to be able to move within and through the structure in order to maintain their activity and function. The most simple cell mimetic surfaces discussed in the literature have been based on transferring lipid bilayers onto glass substrates. Under these conditions, a thin film of water lubricates the interface between the glass and bilayer and allows free lateral diffusion. These types of bilayers generally have poor long-term stability, show limited transmembrane protein activity, and cannot be transferred through the air–water interface without disrupting the structure (46). An alternative approach lies in either the formation of hybrid bilayers, where the inner leaflet is formed from alkane thiol on gold (47), or the deposition of the lipids on a polymer support (46). Hydrogel layers can also be used and act as a hydrated cushion that is both a self-lubricating and a spacer, creating an area for protein insertion without affecting protein function. At present, there are a number of commercial biosensors that utilize this type of technology for disease diagnosis (48).

SURFACE ANALYSIS

The study of surfaces and coatings is an advanced field and ranges from the investigation of elementary chemical processes on single crystals in ultrahigh vacuum (UHV) to the analysis of rather more "dirty" and real surfaces in engineering applications. The development of techniques suitable for surface science has a long history, the driving force for which has only recently included attempting to understand and control biomaterial surface interactions. Table 3 details some of the more common surface analysis techniques used today in the characterization of biomaterials. The physical principles of all the techniques commonly employed today are well understood and have been for many decades. Many of the approaches described here have their origin in the study of elementary chemical and physical processes, the semiconductor industry, and from engineering disciplines.

When considering the choice of surface analytical tools, it is important to appreciate the questions that require answering. A single technique cannot generally provide a complete picture of the surface characteristics. In this section, a description of the some of the most commonly used techniques is provided with reference to more detailed and extensive reviews. It is important to note that the techniques fall into two classes, those that operate inside a vacuum and those that can directly probe the biomaterial–water interface. While it is obviously preferable to use those techniques that can perform under ambient conditions, in general these techniques are either not as informative or not as surface sensitive as the vacuum techniques. For this reason, the vacuum techniques are commonly utilized to provide a detailed characterization, but in doing so it must always be under the assumption that the surface is the same in vacuum as it is under water. This is a rather large assumption, particularly if the material is able to reorganize itself relatively easily. The surface energy change following immersion in water can be rather large, as indicated above, and in mixed biomaterial phases components that are absent at the surface in vacuum may dominate when the material is immersed in an aqueous environment. There is evidence for this kind of surface reorientation in the contact angle hysteresis of water on some polymers.

Hysteresis is the difference in contact angle between a water contact line advancing or receding across a surface. For some polymers, the advancing angle is high and the receding angle is low, indicating that at the polymer–air interface the polymer is hydrophobic and at the polymer–water interface it is hydrophilic. As long as the surface is flat and homogenous, this is evidence of surface reorganization. For some materials it is possible to reduce the rate of reorganization by cooling. This is typically achieved by hydrating the sample in air, and then freezing the sample in liquid nitrogen prior to entry into the vacuum chamber. The sample needs to be held at low temperature while the ice on the surface sublimes and then vacuum techniques

Table 3. Surface Analysis Techniques Used in the Characterization of Biomaterials

Technique	Sampling Depth/Height (Spatial Resolution)	Information Obtained	References
Ultrahigh Vacuum			
Static secondary ion mass spectrometry	<5 nm (500 Å)	Chemical Spectroscopy and Imaging	49,50
X-ray photoelectron spectroscopy	2–10 nm (5 μm)	Elemental, chemical Spectroscopy and Imaging	51
Ambient Techniques			
Attenuated total reflectance Fourier transform Infrared (ATR/FTIR) spectrometry	> 100 nm (1 μm)	Chemical	52
Contact angle measurement	<1 nm (1 mm)	Surface free energy, wettability	53
Atomic force microscopy (AFM) Imaging	Atomic = 20 μm	Topography, coverage, atomic structure	54
Surface force measurement	Å = μm	Chemical, conformational, structural	
Ellipsometry	Å = 300 nm	Layer thickness, adsorption kinetics	55,56
Streaming potential measurements/electroosmosis	Not applicable	Electrokinetics	57
Surface plasmon resonance	~300 nm	Adsorbed mass and adsorption kinetics	58

can be applied to the surface, which should not have reorganized from the hydrated state (59).

Vacuum Techniques

Traditional surface analysis techniques are usually based on ultrahigh vacuum instrumentation. One of the reasons for this was that much of the initial interest in the field was concentrated upon extremely clean and often highly reactive surfaces. To maintain the surface in this state during the analysis it is important to prevent undesired gas or vapor molecules sticking to the surface and changing its characteristics. This is only possible in ultrahigh vacuum ($<10^{-9}$ mbar or so). A second important reason is that to study just the surface and eliminate contributions from the bulk of the material it is necessary to use probe species that strongly interact with matter, such as ions and electrons. These cannot penetrate through more than a few atomic layers, and hence provide highly surface sensitive information. However, the detection of such species normally requires that they travel a considerable distance from the surface. At atmospheric pressure the average distance traveled prior to interaction with gaseous species is too short for the detection systems to work. A vacuum of, typically, 10^{-7} mbar or better is required for the techniques described here to operate. In addition, many of the components necessary for the production and detection of probe species can only be operated in vacuum; at atmospheric pressure, they would be irreparably damaged. We will now briefly describe some of the key surface analysis techniques used in the characterization of biomaterials. More detailed information and discussion on the interpretation of these techniques can be found in books by Vickerman (54) and Watts (51).

X-Ray Photoelectron Spectroscopy

During X-ray photoelectron spectroscopy (XPS) analysis, the sample is illuminated with X rays of a particular energy that causes electrons to be emitted from the sample. This phenomenon is called the photoelectric effect and it is generally found that an X-ray photon imparts all of its energy to a single electron during the process. Since the electrons are bound in orbitals of well-defined energy (binding energy) that are characteristic of the material, the outgoing electrons have a kinetic energy that is essentially the difference between the photon and binding energies. For core level electrons, this binding energy is characteristic of the nucleus to which the electron is closely bound. Only those electrons that are generated close to the surface can escape from the sample without loss of energy due to inelastic collisions with other atoms. Thus, by analyzing the number of electrons emitted from the surface as a function of electron kinetic energy it becomes possible to identify the elements present on, or near to, the material surface. As long as the photon energy is significantly larger than the binding energy of the electron, the probability of generating a photoelectron from a core level is independent of the chemical situation of the element. This means that it is possible not only to identify the elements present but, with appropriate sensitivity factors, quantify the relative amounts of each element.

The chemical situation of the element may, however, have an influence on the binding energy of the core level electron. Chemical bonds to different elements may cause some charge to accumulate on the element of interest, this will directly affect the binding energy of the core electrons. This change in the binding energy is termed the "chemical shift" by analogy to nuclear magnetic resonance (NMR) spectroscopy. The appearance of a chemical shift is extremely useful in surface analysis as it can provide information on how the various elements in the surface are bonded to each other. Where the same element is in a range of chemical environments it is often possible to deduce the fraction of atoms in each environment by careful curve fitting of the spectrum. An example of these types of chemical shifts is illustrated in Fig. 2. In this case, there are chemical shifts evident in a high resolution carbon 1*s*

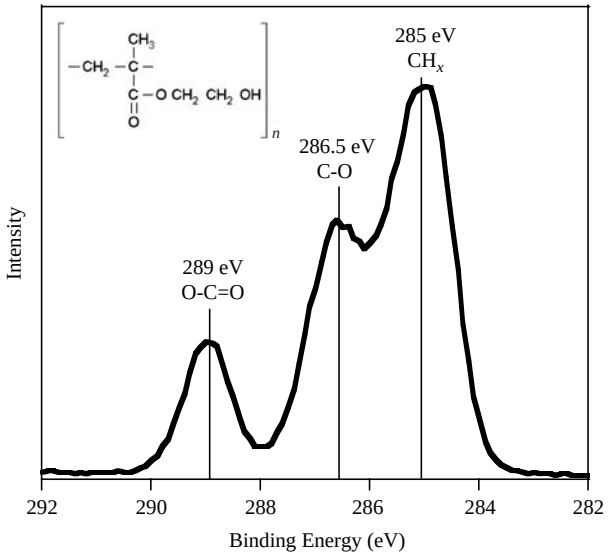


Figure 2. The XPS high resolution C1*s* spectrum of a HEMA contact lens. The figure illustrates the various chemical shifts associated with the chemical bonds in the polymer chain.

(C1*s*) spectrum due to the various chemical bonds present in poly(hydroxyethylmethacrylate), pHEMA, a common soft contact lens material.

The surface sensitivity of XPS is dependent on the angle at which electrons are emitted from the sample. For smooth, flat samples it is possible to enhance the surface sensitivity by analyzing electrons which are emitted at a grazing incidence from the sample. By collecting at a number of different angles it is possible to obtain information on the depth distribution of components close to the material surface. This depth profiling capability is particularly important when thin films and coatings of < 10 nm thickness are being studied.

The XPS has been utilized to chemically characterize biomaterials in four principal areas: identification and characterization of the surface chemistry of bulk polymers; characterization of surface specific modifications; characterization of coatings; detection of biomolecules. Factors that are commonly investigated using XPS include: surface oxidation and reorientation of polymer segments; surface segregation (blooming) of plasticizers, additives, and low molecular weight fragments; adventitious contaminants such as silicones and protein adsorption.

Secondary Ion Mass Spectrometry

The impact of a high kinetic energy (typically 1–100 keV) ion, atom, or molecule causes material to be sputtered from a surface. The origin of the vast majority of ejected species is from the topmost layer of the sample. Therefore analysis of the sputtered fragments can provide information on the composition of the material surface. A small proportion of the ejected atomic and molecular fragments are ionized, these are called the secondary ions. Secondary ion mass spectrometry (SIMS) is the application of mass spectrometry to the secondary ions. Note that SIMS is an ablative, destructive technique and this can be used to advantage in generating a depth profile of layered surfaces. The use of SIMS for depth profiling is termed "dynamic" SIMS and is most often employed in the study of layered materials in which the elemental composition of the layers is of interest, for example, doping levels in semiconductors. It is not possible to obtain more detailed chemical information using dynamic SIMS because of the damage induced by the high energy primary ions. In contrast, "static" SIMS employs a low density, low dose ion bombardment such that the probability of two ion impacts occurring at the same place on the sample is negligibly small ($< 10^{13}$ ions $\cdot$ cm^{-2}). The mass spectrum then contains information that is characteristic of the undamaged surface. This information is particularly useful in the analysis of organic materials, when the normal rules of organic mass spectrometry can be applied to the interpretation of SIMS data. Most modern static SIMS instruments are based on time-of-flight mass analyzers (TOF–SIMS), which have a far greater combined sensitivity and mass resolution than quadrupole or magnetic sector detectors. The probability of ion generation is influenced by a daunting range of factors and thus SIMS is regarded as a nonquantitative technique. However, it is commonly found that in a range of similar materials the characteristic ion intensities are approximately proportional to the concentration of species from which they are generated. With a suitable set of calibration data it is then possible to use SIMS in a quantitative manner. The application of TOF–SIMS in the analysis of biomaterials and biological interfaces has historically revolved around the characterization of polymeric interfaces. This has included the study of degradation pathways for biodegradable polymers, the monitoring of coating chemistries, detection of surface contamination and surface chemical characterization of copolymer systems. The surface sensitivity of TOF–SIMS has lead to its application in the detection and identification of biomolecules adsorbed at interfaces. The process is not without its problems though as the largest ions detected from any protein are the immonium ions ($^{+}NH_2{=}CHR$) from each amino acid (MW<200). As a result of this fragmentation, the identification of proteins is often more like a jigsaw puzzle, where the amino acid fragments have to be pieced together using pattern recognition or multivariate analysis techniques, to identify and quantify the parent molecules (60). These types of statistical analysis are being increasingly used to analyze, compare and reconstruct data collected in TOF–SIMS.

In addition to spectroscopy, TOF–SIMS can be used in an imaging mode to chemically map the surface of a material. There is always a trade off between high special resolution and high mass resolution, but with the advent of liquid metal ion sources (e.g., Ga^+ and In^+), systems are typically capable of spatial resolution of < 10 μm, while retaining atomic mass resolution. As a result there is increasing application of TOF–SIMS for the chemical imaging of a range of biomaterial surfaces. Significantly, developments in ion sources have shown that polyatomic (e.g., Au_3) and cluster ion (C_{60}) sources can significantly improve the molecular ion yield of both biological and polymeric materials. With the development of integrated freeze hydration stages for sample preparation, this has

lead to increased activity in the application of TOF–SIMS in the analysis of cell membranes and other hydrated biological systems (49).

AMBIENT TECHNIQUES

Atomic Force Microscopy

Atomic force microscopy (AFM) can be utilized to characterize surfaces via either an imaging or a spectroscopic mode. There are two common methods of imaging utilized in AFM, contact, and tapping mode. Both can be performed in either air or liquid, a factor that makes AFM particularly attractive in biomaterial research. In contact mode, the tip is scraped across the surface, while in tapping mode, the tip is in intermittent contact with the surface and as such, there is limited substrate disturbance. As a result, tapping mode AFM is more common for the characterization of biomaterials and biological surfaces. Common applications of AFM to biomaterial surfaces include: surface topography and coating continuity assessment, measurement and monitoring of coating thickness (see Fig. 3) and phase imaging. The last application is an extension of tapping mode imaging that gives nanometer-scale two-dimensional (2D) information about surface structure. It can be used to locate and characterize the distribution of discrete phases within polymer blends such as polyurethane–urea, and a range of other polymers of biological importance. There is also considerable interest in using AFM to image the surfaces of cells and biomolecules on surfaces (61,62).

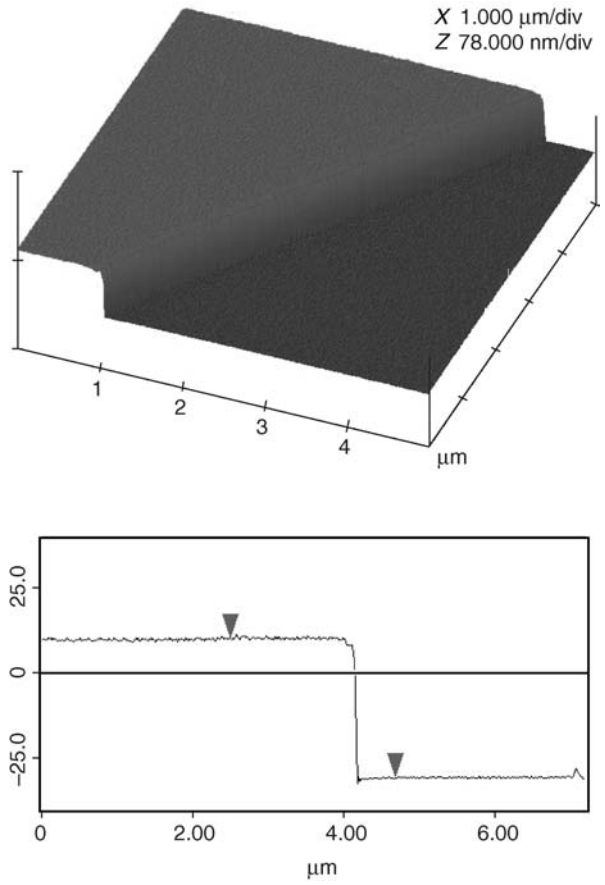


Figure 3. The AFM tapping mode image of a 40 nm step between coated and uncoated regions on a silicon wafer. The step is produced by masking a section of the sample surface prior to plasma polymerization. Once the coating is deposited, the mask is removed and surface imaged. This enables the plasma polymer thickness to be measured with nm resolution. (Image courtesy of Dr. P.G. Hartley, CSIRO Molecular Science, Australia.)

Surface Force Measurements

The forces that act between particles and surfaces determine a wide range of interactions. They control the stability of dispersions and emulsions; determine the adhesion of colloids onto surfaces and the adsorption properties of proteins and cells at surfaces. Surface force measurements are used in variety of industries and increasingly there is interest in the application of surface force analysis to biomaterials. with the aim of characterizing protein–protein, cell–protein, protein–surface and cell–surface interactions.

There are a number of techniques that can be used to measure force interactions between surfaces. They are divided into two classes based on the method used to determine surface separation. Absolute surface separation can only be determined using an interferometric technique such as the surface force apparatus (SFA). These techniques are limited by the need for a transparent substrate and specific geometric configuration. In response, noninterferometric techniques have been developed that can employ a wider variety of substrates, and that rely on indirect determination of the surface separation rather than interferometry. One of these is AFM surface force measurement.

Force measurements can be made with an AFM using both a bare tip and a tip modified with a probe particle. If the results from these measurements are to be quantitative, knowledge of the radius of curvature for the probe or tip is critical. While it is possible to measure the nominal radius of a bare AFM tip, the indirect nature of the measurement adds to the error associated with the resulting force calculations (63). The Derjaguin approximation is only valid when the radius is much larger than the surface separation, which is not necessarily the case if an unmodified tip is used. If a colloid probe is utilized, the radius of curvature can be measured easily and accurately using scanning electron microscope (SEM) or optical microscopy.

The zero surface separation for noninterferometric techniques are set at the hard wall encountered when the two surfaces are forced together. Termed the constant compliance regime, this is the region of the force curve in which the displacement of the colloid probe is linear with respect to the surface motion. This assumption of hard wall contact is an inherent limitation of these techniques, particularly if the surface deforms or compresses under pressure, as is often the case with polymer surfaces. The compression of a polymer layer can have a number of effects on the force curve. In the first instance, compression of a dense polymer layer can result in a compliance line mimicking hard wall contact, with a layer of compressed polymer between the probe and the substrate. If the polymer is less densely

packed, the probe may displace the material, squeezing the polymer out of the gap between the probe and the surface. This results in discontinuities in the compliance region as the force increases and the probe pushes through the polymer layers (63). As a result, conclusions about absolute layer thickness cannot be inferred from this data.

While XPS is able to characterize the chemistry of a surface in the dry state, surface force measurements are well suited to characterizing the intermolecular forces and stability in a variety of environments. A number of studies have investigated the surface force characteristics of rfgd films, grafted polymer layers, and adsorbed protein films in a variety of media (64). Other studies have used surface force measurements to characterize the interactions of polymer modified surfaces in an attempt to elucidate parameters that control the structure of the polymer layer (64). Surface force techniques have been used to investigate the effects of molecular weight, ionic strength, charge density, and polymer concentration on the interaction forces of adsorbed and grafted polymers layers. Increasingly, strategies to eliminate protein adsorption are based on the characterization and modification of the surfaces and thus the interaction forces that govern protein adsorption (65). A number of theoretical studies have also used surface force interactions as design parameters when modeling polymer coatings capable of resisting protein adsorption (66).

In addition to these standard modes of operation, by chemically modifying the AFM cantilever it is possible to map specific interactions between the tip and the surface. Depending on the type of modification made to the cantilever, a range of interactions can be investigated. With a cantilever modified with a receptor specific integrin, dynamic force spectroscopy can be used to identify and map receptor sites on a cell surface (67). By modifying the cantilever with specific chemical functional groups, differences in frictional properties and the distribution of different phases can be probed where there is no topographical variation (68).

Optical Techniques

The refractive index close to an interface can be measured by a number of optical techniques. The two most commonly employed for this purpose are surface plasmon resonance (SPR) (58) and ellipsometry (55,56). Since proteins have a higher refractive index ($\sim$1.55) than water ($\sim$1.33) it is possible to monitor the amount of protein adsorption at an interface through a measurement of the refractive index and thickness of the adsorbed layer. These techniques have found utility in a wide range of areas relevant to biomaterials research, such as the study of protein adsorption to biomaterials. ligand–receptor interactions and the dissolution and swelling of polymers. The advantage over traditional approaches such as enzyme linked immunosorbent assay (ELISA), fluorescent labeling or radiolabeling is that the proteins under study do not have to be chemically altered in any way.

The disadavantages of these techniques are that the substrate must be flat and conform to a number of optical requirements for the techniques. Additionally, these techniques cannot directly distinguish between different proteins, since all proteins have roughly equivalent optical densities. To determine the identity of proteins adsorbed from a mixed solution it is necessary to subsequently expose the surface to antibodies specific to each protein of interest. Binding of the antibody can be monitored as an increase in the adsorbed layer thickness, however, this approach is difficult to employ quantitatively as there may be nonspecific and competitive adsorption of the antibody as well as a limited availability of binding sites on the adsorbed target protein.

Surface Plasmon Resonance

A surface plasmon is a collective oscillation of electrons that can be excited in certain metals such as silver and gold. The frequency of this oscillation depends on the refractive index of the dielectric material close to the metal interface. If the metal is a thin film it is possible to excite surface plasmons by reflecting light of a wavelength greater than the thickness of the metal from the reverse side of the film. The ability to cause this excitation depends on the wavelength of light, the refractive index of the material through which the light travels (which remains constant in this geometry) and the angle of reflection. When the conditions are correct, light is absorbed. In the usual set up for surface plasmon resonance (SPR) instruments the light undergoes total internal reflection at the interface and the angle at which light is absorbed is monitored. If there is a change in the refractive index close to the interface then a corresponding shift in the angle at which light is absorbed can be followed. The sensitivity of SPR decreases exponentially in distance from the surface with a decay length of the order of the wavelength of the light. If the layer to be analysed is significantly smaller than the decay length, which is usual for protein adsorption, then it can be assumed that any change in refractive index is proportional to the mass of adsorbed protein.

Ellipsometry

When light is reflected at an oblique angle from a planar surface it commonly undergoes a change in polarization. By analyzing these changes, it is possible to infer both the optical properties and thickness of thin films on the surface. To obtain the most complete characterization, a large number of different wavelengths of light or different angles of incidence must be studied. The measured data is then compared to the expected polarization changes calculated from a model, and parameters in the model changed (such as thickness or refractive index) to find a fit between the two. The sensitivity of ellipsometry is comparable to SPR ($\sim 0.01/\mathrm{g{\cdot}cm^2}$), however, it is able to analyze comparatively thick layers of material.

CONCLUSION

While materials selection for most biomedical devices needs to be based upon bulk properties, in this article we have provided a broad overview of the basic properties of surfaces, and introduced some of the reasons why the surface properties may significantly 3 influence the efficacy of biomaterials

and biomedical devices. Surface modification aims to tailor the surface characteristics of a material for a specific application without detrimentally affecting the bulk properties. Throughout this article we have shown how a range of physical, chemical, and biological modifications can be made to surfaces and used to manipulate surface characteristics. Finally, we discussed a range of highly sensitive surface analytical methods that can be utilized to investigate both the nature of an interface and its interactions with biological environments. As is always the case with review articles of this type, it is impossible to give detailed accounts of all of the material being discussed. We have included a range of references (*Reading List*) to aid the reader in further developing their understanding of each of the specific concepts and techniques. Additionally, we have included a list of more general references that cover many of the fundamental concepts discussed within this article.

BIBLIOGRAPHY

Cited References

1. Mittal KL, editor. Contact Angle, Wettability and Adhesion. Utrecht: VSP; 1993.
2. Berry CC, Campbell G, Spadiccino A, Robertson M, Curtis ASG. The influence of microscale topography on fibroblast attachment and motility. Biomaterials 2004;25(26):5781–5788.
3. Andrade JD. Principles of Protein Adsorption. In: Andrade JD, editor. Surfaces and Interfacial Aspects of Biomedical Polymers. Vol. 2: Protein Adsorption. New York: Plenum Press; 1985.
4. Underwood PA, Steele JG. Practical limitations of estimation of protein adsorption to polymer surfaces. J Immunol Methods 1991;142(1):83–94.
5. Leduc CA, Vroman L, Leonard EF. A Mathematical-Model for the Vroman Effect. Ind Eng Chem Res 1995;34(10):3488–3495.
6. Hubbell JA. Bioactive Biomaterials. Curr Opin Biotechnol 1999;10:123–129.
7. Bray D. Cell Movements:From Molecules to Motility. 2nd ed. New York: Garland; 2001. p 372.
8. Chu PK, Chen JY, Wang LP, Huang N. Plasma-surface modification of biomaterials. Mat Sci Eng R 2002;36(5-6): 143–206.
9. Favia P, d'Agostino R. Plasma Treatments and Plasma Depositions of Polymers for Biomedical Applications. Surf Coat Tech 1998;98:1102–1106.
10. Aronsson BO, Lausmaa J, Kasemo B. Glow discharge plasma treatment for surface cleaning and modification of metallic biomaterials. J Biomed Mater Res 1997;35(1):49–73.
11. Mandl S, Rauschenbach B. Plasma immersion ion implantation. New technology for homogeneous modification of the surface of medical implants of complex shapes. Biomed Tech 2000;45(7–8):193–198.
12. Bilek MMM, McKenzie DR, Tarrant RN, Lim SHM, McCulloch DG. Plasma-based ion implantation utilising a cathodic arc plasma. Surf Coat Tech 2002;156(1–3):136–142.
13. Shin GH, Lee YH, Lee JS, Kim YS, Choi WS, Park HJ. Preparation of plastic and biopolymer multilayer films by plasma source ion implantation. J Agric Food Chem 2002;50(16):4608–4614.
14. McPherson TB, Shim HS, Park K. Grafting of PEO to glass, nitinol, and pyrolytic carbon surfaces by gamma irradiation. J Biomed Mater Res 1997;38(4):289–302.
15. Benson RS. Use of radiation in biomaterials science. Nucl Instrum Meth B 2002;191:752–757.
16. Welle A, Gottwald E. UV-based patterning of polymeric substrates for cell culture applications. Biomed Microdevices 2002;4(1):33–41.
17. Zhang F, Kang ET, Neoh KG, Wang P, Tan KL. Surface modification of stainless steel by grafting of poly(ethylene glycol) for reduction in protein adsorption. Biomaterials 2001;22(12):1541–1548.
18. Krupa D, Baszkiewicz J, Kozubowski J, Barcz A, Sobczak J, Bilinski A, Rajchel B. The influence of calcium and/orphosphorus ion implantation on the structure and corrosion resistance of titanium. Vacuum 2001;63(4): 715–719.
19. Braceras I, Alava JI, Onate JI, Brizuela M, Garcia-Luis A, Garagorri N, Viviente JL, de Maeztu MA. Improved osseointegration in ion implantationtreated dental implants. Surf Coat Tech 2002;158:28–32.
20. Cui FZ, Luo ZS. Biomaterials modification by ion-beam processing. Surf Coat Tech 1999;112(1–3):278–285.
21. Bambauer R, Mestres P, Schiel R, Schneidewind JM, Latza R, Bambauer S, Sioshansi P. Surface treated catheters with ion beam based process for blood access. Ther Apher 2000;4(5):342–347.
22. Li DJ, Zhao J, Gu HQ. Hemocompatibility of DLC coatings synthesized by ion beam assisted deposition. Sci China Ser E-Technol Sci 2001;44(4):427–431.
23. Kingshott P, Griesser HJ. Surfaces that resist bioadhesion. Curr Opin Solid St M 1999;4:403–412.
24. Bures P, Huang YB, Oral E, Peppas NA. Surface modifications and molecular imprinting of polymers in medical and pharmaceutical applications. J Control Release 2001;72(1–3):25–33.
25. Kato K, Uchida E, Kang ET, Uyama Y,Ikada Y. Polymer surface with graft chains. Prog Polym Sci 2003;28(2):209–259.
26. Cai KY, Lin SB, Yao KD. Advances in research on surface engineering of biomaterials for tissue engineering. Prog Chem 2001;13(1):56–64.
27. Sakiyama-Elbert SE, Hubbell JA. Functional biomaterials: Design of novel biomaterials. Ann Rev Mater Res 2001; 31:183–201.
28. Massia SP, Stark J. Immobilized RGD peptides on surface-grafted dextran promote biospecific cell attachment. J Biomed Mater Res 2001;56(3):390–399.
29. Whitesides GM, Ostuni E, Takayama S, Jiang X, Ingber DE. Soft Lithography in Biology and Biochemistry. Annu Rev Biomed Eng 2001;3:335–373.
30. Beck AJ, Jones FR, Short RD. Plasma copolymerization as a specific route to the fabrication of new surfaces with controlled amounts of specific chemical functionality. Polymer 1996;37:5537–5539.
31. Shen MC, Martinson L, Wagner MS, Castner DG, Ratner BD, Horbett TA. PEO-like plasma polymerized tetraglyme surface interactions with leukocytes and proteins: in vitro and in vivo studies. J Biomater Sci Polym Ed 2002;13(4):367–390.
32. Pan YV, Wesley RA, Luginbuhl R, Denton DD, Ratner BD. Plasma polymerized *N*-isopropylacrylamide: synthesis and characterization of a smart thermally responsive coating. Biomacromolecules 2001;2(1):32–36.
33. Fraser S, Short RD, Barton D, Bradley JW. A multi-technique investigation of the pulsed plasma and plasma polymers of acrylic acid: Millisecond pulse regime. J Phys Chem B 2002;106(22):5596–5603.
34. Han LCM, Timmons RB. Pulsed-plasma polymerization of 1-vinyl-2- pyrrolidone: Synthesis of a linear polymer. J Polym Sci Pol Chem 1998;36(17):3121–3129.

35. Hanawa T. In vivo metallic biomaterials and surface modification. Mat Sci Eng A-Struct 1999;267(2):260–266.
36. Guzman L, Celva R, Miotello A, Voltolini E, Ferrari F, Adami M. Polymer surface modification by ion implantation and reactive deposition of transparent films. Surf Coat Tech 1998;104:375–379.
37. Cui FZ, Luo QL, Feng J. Highly adhesive hydroxyapatite coatings on titanium alloy formed by ion beam assisted deposition. J Mater Sci Mater M 1997;8:403–405.
38. Howlett CR, Zreiqat H, Wu Y, McFall DW, McKenzie DR. Effect of ion modification of commonly used orthopedic materials on the attachment of human bone-derived cells. J Biomed Mater Res 1999;45(4):345–354.
39. Sioshansi P, Tobin EJ. Surface treatment of biomaterials by ion beam processes. Surf Coat Tech 1996;83(1–3):175–182.
40. Davenas J, Thevenard P, Philippe F, Arnaud MN. Surface implantation treatments to prevent infection complications in short term devices. Biomol Eng 2002;19(2–6):263–268.
41. Leng YX, Chen JY, Zeng ZM, Tian XB, Yang P, Huang N, Zhou ZR, Chu PK. Properties of titanium oxide biomaterials synthesized by titanium plasma immersion ion implantation and reactive ion oxidation. Thin Solid Films 2000;377:573–577.
42. Ostuni E, Chapman RG, Holmlin RE, Takayama S, Whitesides GM. A survey of structure-property relationships of surfaces that resist the adsorption of protein. Langmuir 2001;17:5605–5620.
43. Textor M, Ruiz L, Hofer R, Rossi A, Feldman K, Hahner G, Spencer ND. Structural chemistry of self-assembled monolayers of octadecylphosphoric acid on tantalum oxide surfaces. Langmuir 2000;16(7):3257–3271.
44. Chandy T, Das GS, Wilson RF, Rao GHR. Use of plasma glow for surface-engineering biomolecules to enhance bloodcompatibility of Dacron and PTFE vascular prosthesis. Biomaterials 2000;21(7):699–712.
45. Peluso P, Wilson DS, Do D, Tran H, Venkatasubbaiah M, Quincy D, Heidecker B, Poindexter K, Tolani N, Phelan M, Witte K, Jung LS, Wagner P, Nock S. Optimising antibody Immobilization strategies for the construction of protein microarrays. Anal Biochem 2003;312:113–124.
46. Sackmann E, Tanaka M. Supported membranes on soft polymer cushions: fabrication, characterization and applications. Trends Biotechnol 2000;18:58–64.
47. Plant AL. Supported hybrid bilayer membranes as rugged cell membrane mimics. Langmuir 1999;15(15):5128–5135.
48. Krishna G, Schulte J, Cornell BA, Pace RJ, Osman PD. Tethered bilayer membranes containing ionic reservoirs: Selectivity and conductance. Langmuir 2003;19(6):2294–2305.
49. Winograd N. Prospects or imaging TOF-SIMS: from fundamentals to biotechnology. Appl Surf Sci 2003;203:13–19.
50. Castner DG, Ratner BD. Biomedical surface science: Foundations to frontiers. Surface Sci 2002;500(1–3):28–60.
51. Watts JF, Wolstenholme J. An Introduction to Surface Analysis by XPS and AES. Chichester: John Wiley & Sons; 2003.
52. Chittur K. FTIR/ATR for protein adsorption to biomaterials surfaces. Biomaterials 1998;19:357–369.
53. Adamson AP, Gast AW. Physical Chemistry of Surfaces. New York: John Wiley & Sons; 1997.
54. Vickerman JC, editor. Surface analysis: The Principal Techniques. Chichester: John Wiley & Sons; 1997, p. 457.
55. Elwing H. Protein adsorption and ellipsometry in biomaterials research. Biomaterials 1998;19:397–406.
56. Arwin H. Ellipsometry on thin organic layers of biological interest: characterization and applications. Thin Solid Films 2000;377:48–56.
57. Hunter RJ. Zeta Potential in Colloid Science. London: Academic Press; 1988. p 67.
58. Green RJ, Frazier RA, Shakesheff KM, Davies MC, Roberts CJ, Tendler SJB. Surface plasmon resonance analysis of dynamic biological interactions with biomaterials. Biomaterials 2000;21(18):1823–1835.
59. Lewis KB, Ratner BD. Observation of surface restructuring of polymers using ESCA. J Colloid Interface Sci 1993;159:77–85.
60. Wagner MS, Castner DG. Characterization of adsorbed protein films by time of flight secondary ion mass spectrometry (ToF-SIMS) in conjunction with prinicpal component analysis (PCA). Langmuir 2001;17:4649–4660.
61. Boonaert C, Rouxhet P, Dufrene Y. Surface propaerties of microbial cells probed at the nanometre scale with atomic force microscopy. Surf Interface Anal 2000;30:32–35.
62. Fritz M, Radmacher M, Cleveland JP, Allersma MW, Stewart RJ, Gieselmann R, Janmey P, Schmidt CF, Hansma PK. Imaging globular and filamentous proteins in physiological buffer solutions with tapping mode atomic force microscopy. Langmuir 1995;11:3529–3535.
63. Hartley PG. Farinato R, Dubin P, editors. Measurement of Colloidal Interactions Using the Atomic Force Microscope, in Colloid-Polymer Interactions: From Fundamentals to Practice. New York: John Wiley & Sons; 1999.
64. Hartle PG, McArthur SL, McLean KM, Griesser HJ. Physicochemical properties of polysaccharide coatings based on grafted multilayer assemblies. Langmuir 2002;18(7):2483–2494.
65. Leckband D, Sheth S, Halperin A. Grafted poly(ethylene oxide) brushes as nonfouling surface coatings. J Biomater Sci Polym Ed 1999;10(10):1125–47.
66. Halperin A. Polymer brushes that resist adsorption of model proteins: Design parameters. Langmuir 1999;15:2525–2533.
67. Evans E. Energy landscapes of biomolecular adhesion and receptor anchoring at interfaces explored with dynamic force microscopy. Faraday Discuss 1998;111:1–16.
68. Sun S, Chong KS, Leggett GJ. Nanoscale molecular patterns fabricated by using scanning near-field optical lithography. J Am Chem Soc 2002;124(11):2414–2415.

Reading List

Andrade JD, editor. Surfaces and Interfacial Aspects of Biomedical Polymers. New York: Plenum Press; 1985.
Malmsten M, editor. Biopolymers at Interfaces. New York: Marcel Dekker; 1998/2004.
Castner DG, Ratner BD. Biomedical surface science: Foundations to frontiers. Surface Sci 2002;500(1–3):28–60.
Adamson AW, Gast AP. Physical Chemistry of Surfaces. New York: John Wiley & Sons; 1997.

See also BIOCOMPATIBILITY OF MATERIALS; BIOSURFACE ENGINEERING; MICROSCOPY, SCANNING TUNNELING.

BIOMATERIALS, TESTING AND STRUCTURAL PROPERTIES OF

DONGLU SHI
University of Cincinnati
XUEJUN WEN
Clemson University
Clemson, South Carolina

INTRODUCTION

Tissue transplantation and synthetic devices have been utilized in order to substitute the function of lost

or damaged hard tissue, such as bone and tooth. Tissue transplants can be autologous, allogeneic, or xenogeneic. However, the use of autologous tissue involves additional surgery and donor site morbidity while the use of allogeneic or xenogeneic tissue involves the risks of immune rejection and disease transmission. Therefore, synthetic hard tissue implants are very necessary. Metals, ceramics, composites, and even polymers are investigated as candidates for the hard tissue replacements. For heavy loaded applications, such as hip prostheses, metals (e.g., Ti–alloys, Co–Cr), and strong inert ceramics (e.g., alumina, zirconia) are extensively studied. Unfortunately, various problems related to both the metallic materials and the bioinert ceramics, for example, corrosion, elastic modulus mismatch (stress concentration and shielding), and bioinertness (only physical connection with host) with metals, and brittle, elastic modulus mismatch, and bioinertness with bioinert ceramics. For these reasons, bioceramics are showing very promising results in the high bioactivity and the formation of interfacial chemical bond with host tissue, which was called osseointegration (1). So far, several bioactive ceramics have been proposed for hard tissue replacements, hydroxyapatide (HA) and bioactive glasses are the most acceptable materials for hard tissue applications (2). The advantages of bioceramics over inert ceramics and metals allow for developing better hard tissue replacements with the characteristics of bioactive and elastic modulus more close to that of bone (2). On the other hand, the mechanical properties of bioceramics are fairly poor when compared with their replaced natural hard tissues. The poor mechanical properties, especially inside the body aqueous environments, limit their applications to only small, unloaded, and low loaded implants, powders, coatings, composites, porous scaffolds for tissue engineering, and so on. Bioceramic coatings and porous scaffold are showing the most promising results for the future hard tissue replacements (3–5). There are various methods developed to produce HA coatings (3–5). Among these techniques, plasma spraying has widely been used. However, this method is not applicable for deposition HA films onto a porous substrate.

In order to obtain a bone substitute possessing both desirable mechanical properties and bioactivity, two major deposition routes in coating the bioactive HA on a highly porous alumina substrate with the similar range of tensile and compressive strength as natural bone were developed. Coated reticulated bioactive substrates can provide the needed mechanical strength for the replacement of the bear-loading functions. The first one is a suspension method in which the ceramic substrates are first coated with a suspension containing the HA powder followed by a sintering with an appropriate time–temperature cycle to densify the HA coating. The second one is a synthesis route or called thermal deposition.

The techniques of coating uniformly thin layers of bioactive HA onto highly porous alumina substrate, the structural properties, especially the interfaces between the coating and the substrate, and the bioactive behavior of the coated substrate in the simulated body fluid (SBF) will be presented in this article.

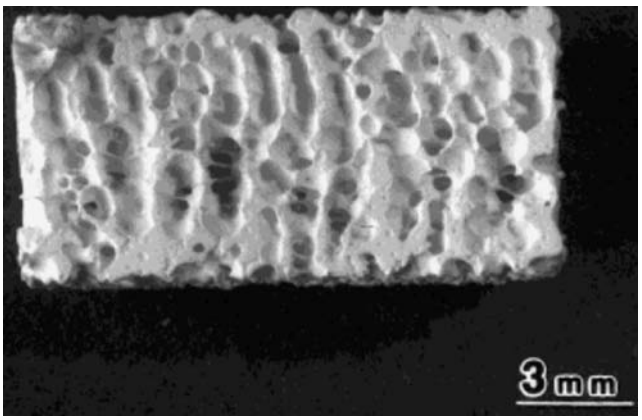


Figure 1. Cross-section of reticulated alumina substrate, showing the interconnected porosity.

MATERIALS AND METHODS

Reticulated alumina Al_2O_3 was used as the substrate materials. Figure 1 is the gross morphology of the substrate; and Fig. 2 is the cross-section showing the interconnected pores. The average size of the pores is 500 μm, which are large enough to allow the ingrowth of bone tissue. Substrates were cleaned using ultrasonic cleaner in acetone and dried at 100 °C before applying the coating.

Suspension Method

The coating suspension was made up of finely milled ceramic powders, an organic solvent, and a binder. The binder was used to prevent the precipitation of particles and to provide bonding strength to the coating after drying. One important property of the suspension is its viscosity. Specifically, when the porous substrate was immersed in the coating suspension, the suspension must be fluid enough to enter, fill, and uniformly coat the substrate skeleton. Low viscosity could result in undesirable thin films while highly viscous slurry would block the pore, thus impairing the interconnectivity of the pores. The viscosity

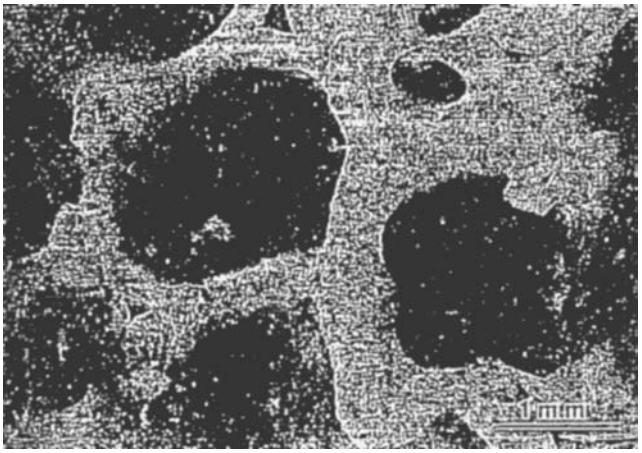


Figure 2. Cross-section of reticulated alumina substrate, showing the interconnected porosity.

Table 1. Sample Group Assignment

Sample	Source and Treatment	SSA, $m^2 \cdot g^{-1}$
HA	Commercial	63.02
HA600	HA heated at 600 °C, 30 min	40.92
HA900	HA heated at 900 °C, 30 min	17.45
SHA700	Synthesized HA heated at 700 °C, 4 h	15.19
SHA800	Synthesized HA heated at 800 °C, 4 h	1.49
SHA900	Synthesized HA heated at 900 °C, 4 h	1.23
No. 1	HA heated at 700 °C, 3.5 h, particle size: 40–100 mesh	24.38
No. 2	HA heated at 700 °C, 3.5 h, particle size: 100–200 mesh	25.70
No. 3	HA heated at 700 °C, 3.5 h, particle size: $>$ 200 mesh	27.02

was controlled by the relative amount of solvent, binder, and particles. The porous substrates were subsequently immersed into the mixture. After the porous substrates were completely infiltrated, they were spun briefly in a high speed centrifuge for removal of excess solution. Coated substrates were dried in an oven at 100 °C. The dried specimens were heated in air to 400 °C for 1 h to burn out the organic binder from suspension. During the burnout process, a slow and controlled heating rate was necessary to avoid bubbling in the coating. Then the samples were subject to sintering at different temperatures. Two types of suspension were developed for coating. One was prepared by suspending HA particles (300 mesh) (Chemat Technology, Inc.) in an organic binder–solvent system without glass sintering aid, glass frits. The other was prepared by partially substitution of HA by sintering aid, glass frits (65%), which have good adhesion to the Al_2O_3 substrate and a weak reaction with HA during firing. The glass frits used in this work were borosilicate glasses containing $\sim$75% of a mixture of SiO_2 and B_2O_3, and 20 wt% alkali metal oxides. After melting, the glass was quenched in water and ground in a ball mill into a glass frit of the desired particle size (325 mesh).

Thermal Deposition

Mixing calcium 2-ethyl hexanoate with bis(2-ethyhexyl) phosphite stoichiometrically in ethanol. The viscosity of the solution was controlled by the quantity of ethanol added. The mixture was stirred for 2 days at room temperature. Then mixture was used to coat porous substrates. The coating method used is the same as used in suspension method described earlier. Coated substrates were air-dried and calcinated up to 1000 °C at a heating rate of 2 °C/min. Then the samples were subject to sintering at different temperatures. For phase analysis purpose, the HA was prepared in the powder form as well through same procedures. Briefly, the solution was open to the air and stirred to vaporize the solvent in chemical hood. Finally, a highly viscous, translucent mixture was obtained and then subject to calcinations at desirable temperatures.

Mechanical strength measurements were carried out on an Instron testing unit. Bars of the porous substrate ($5\times5\times60$ mm^3) were cut using a diamond saw. Tension tests were performed in three-point bending. Compression tests were made on cylinder-shaped sample of 10-mm height and 23 mm diameter. X-ray diffraction (XRD), Scanning electron microscopy (SEM) with an energy dispersive spectrometer (EDS) was utilized to study the coating structure, surface morphology, and the interface structure. The coating bonding strength was measured through tape test (ASTM D 3359), which was originally designed for organic coatings on metallic substrates. This method was used to find the relative bonding strength. All the tests were performed on dense alumina substrates with one or multilayer HA coatings. Permacel 670 tape (Permacel) was used in the test. After removal from the coating, the tape was examined under a light microscope. Sintering process and chemical bonding of the sintering products were examined using differential thermal analysis (DTA) and Fourier transform infrared spectroscopy (FTIR).

The *in vitro* tests were conducted to evaluate the bioactivity of the synthetic HA produced by thermal deposition method (commercial HA as control). All the samples were tested in the powder form (Table 1). Table 1 summarizes the different treatments used to obtain a variety of crystalline structures in the materials; and also names the samples according to the treatment conditions, such as HA, HA600, HA900, SHA700, SHA800, SHA900, No. 1, No. 2, and No. 3. The HA sample group refers to commercial hydroxyapatite samples. HA600 and HA 900 groups refer to commercial HA samples treated for 30 min under 600 and 900 °C, respectively. The SHA700, SHA800, SHA900 conditions refer to synthesized HA using the thermal deposition method described earlier and heated for 4 h at 700, 800, and 900 °C, respectively. Sample No. 1, No. 2, and No. 3 refer to commercial HA and are heat-treated at 700 °C for 3.5 h and with different specific surface area (SSA). Sample No. 3 has the highest SSA; Sample No. 1 has the lowest SSA; and Sample No. 2 is in the middle. The simulated body fluid (SBF) solution that had ionic concentrations close to human blood plasma, as shown in Table 2, was prepared by dissolving reagent grade NaCl, $NaHCO_3$, KCl, $K_2HPO_4{\cdot}3H_2O$, $MgCl_2{\cdot}3H_2O$, $CaCl_2$, and Na_2SO_4 in ion-exchanged distilled water. The solution was buffered at pH 7.4 with 1 *M* HCl and tris(hydroxymethyl) aminomethane, $(CH_2OH)_3CNH_2$ at 37 °C. Powders were immersed into solution at 1 mg/mL ratio and maintained at 37 °C at periods ranging from 15 min to 9 weeks. The calcium concentrations in the solutions were measured by inductively coupled plasma (ICP). Subsequent to immersion, the solutions were vacuum filtered. The powders were gently rinsed with alcohol, ion-exchanged distilled water and then dried at room temperature. The surface microstructures before and after immersion of SBF solution were analyzed via scanning electron microscope (SEM). Both

Table 2. Ionic Concentration of SBF in Comparison With Those of Human Blood Plasma

	Concentration, m*M*							
	Na^+	K^+	Ca^{2+}	Mg^{2+}	HCO_3^-	Cl^-	HPO_4^{2-}	SO_4^-
Blood plasma	142.0	5.0	2.5	1.5	27.0	103.0	1.0	0.5
SBF	142.0	5.0	2.5	1.5	4.2	147.8	1.0	0.5

XRD and FTIR determined the contents of the phases that were present in the coatings. Measurements were obtained on a Philips X-ray diffractometer with CuK-radiation at 35 kV and 23 mA.

RESULTS

Suspension Method

The phase diagram for anhydrous calcium phosphates (Fig. 3) shows that the liquid phase appears at a temperature $> 1500\,^{\circ}C$, and presumably the induced liquid could improve the bonding between HA and the Al_2O_3 substrate during sintering. However, such as liquid-enhanced bonding process was not experimentally observed. Meanwhile, XRD analysis of the coating made from HA solution (without glass additive) showed that HA was decomposed, and in turn, the bioactivity of the coating was changed. A high density of cracks was found to exist in the coating. The adherence of the coating to the Al_2O_3 substrate was low and the coating layer could be peeled off by scraping. These results indicated that HA solution without sintering aid, such as glass frits, was not suitable for this particular application.

Using the sintering aid, glass frits, a well-bonded HA coating was produced. Figures 4–6 are SEM photographs of the surfaces and interfaces of the coatings made from solution with glass frits. As is apparent from Figs. 5 and 6, the glass–HA ceramic layer is firmly attached to the alumina substrate. An average coating thickness is 15 μm. The interfacial strength between the coating and the

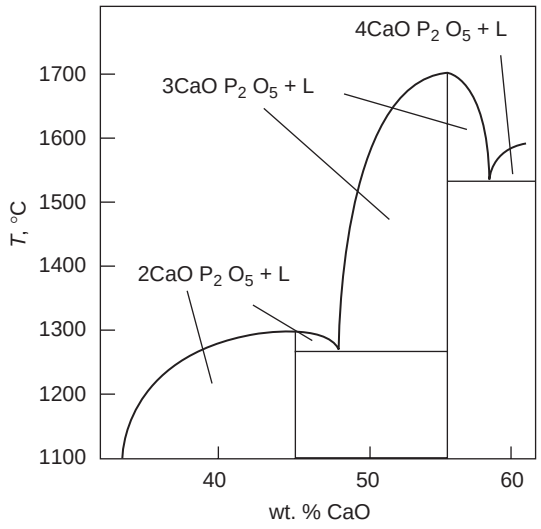


Figure 3. Phase diagram of the system $CaO–P_2O_5$, indicating the appearance of the liquid phase at high temperatures.

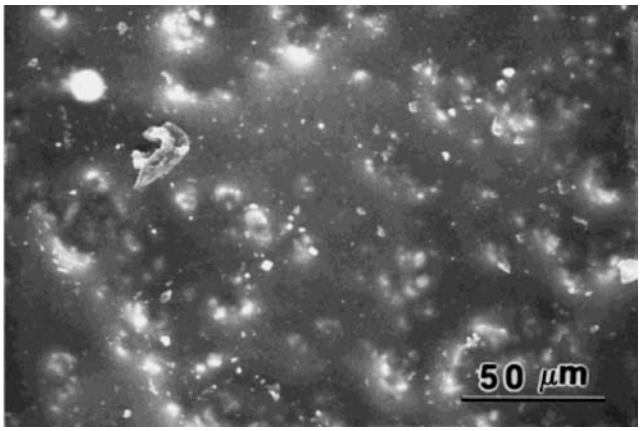


Figure 4. The SEM photograph of the coated surface.

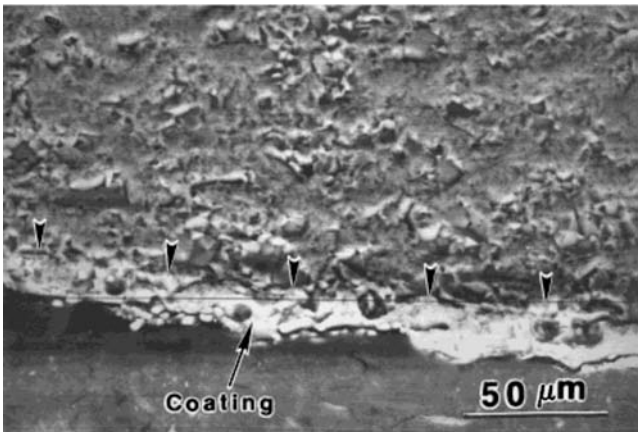


Figure 5. The SEM photograph of the coating interface for dense alumina. The interface is indicated by arrows.

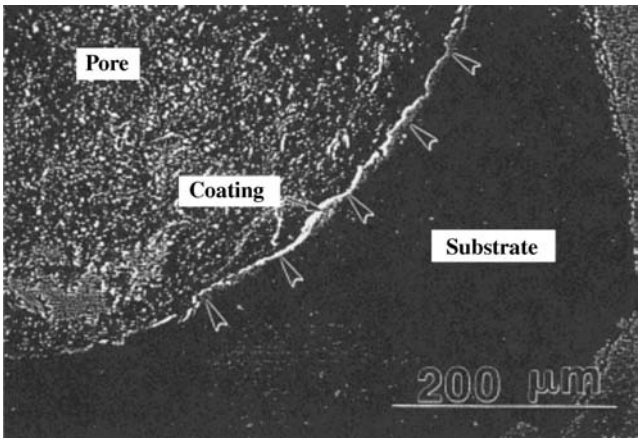


Figure 6. The SEM photograph of the coating interface for porous alumina. The interface is indicated by arrows.

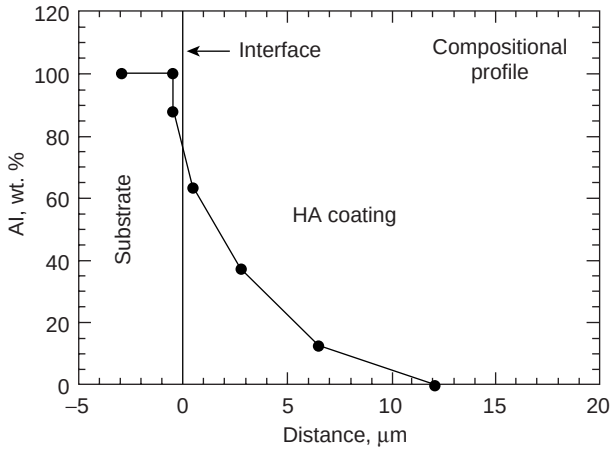


Figure 7. Aluminum content profile along the interface. The solid line is a guide to the eye.

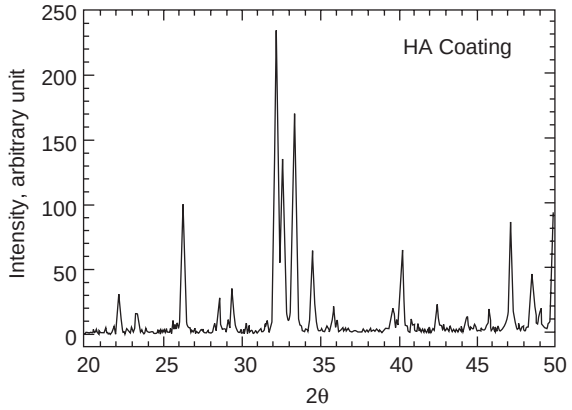


Figure 8. The XRD spectra showing a typical HA pattern of the coating surface.

substrate depends on the adherence of the glass to the Al_2O_3. Figure 7 is the aluminum compositional profile across the interface. The data showed that the aluminum concentration decreased when scanning from the substrate to the outer surface of the coating. It was concluded that aluminum ions diffused during sintering, and consequently bonded the glass to the alumina substrate by ion diffusion. This diffusion bonding is attributed to the formation of a eutectic compound at the interface during the sintering, and thus ensures the strong bonding between the coating and the substrate.

The above results indicate that the development of glass frits is essential for having an excellent adhesion of HA coating to the alumina substrate. Great efforts were therefore made in the preparation of the glass frits. As a sintering aid, the glass must wet the substrate and HA, and its melting point should be lower than the decomposing temperature of HA (1300 °C). Furthermore, for successful coating, optimizing the coefficient of thermal expansion of the glass to match the substrate is critical. It has been known that the mismatch in expansion coefficients between the coating and the substrate materials will give rise to interfacial stress that weakens the bonding strength or leads to the cracking and spalling of the coating. The magnitude of this stress is proportional to the difference between the thermal expansion coefficients of the coating and the substrate. The expansion coefficient of HA (13.3×10^{-6}/°C) is relatively higher than that of alumina (8.0×10^{-6}/°C). The expansion coefficient of the selected glass should then be an intermediate one to reduce this difference. Another important aspect of the glass is chemical durability. For biological applications, it is essential for glass to be nontoxic and stable in the body fluid. The dissolution of the glass will lead to the degradation of the coating. The HA particles will escape from the coating, and this will have an extremely negative effect, such as interfacial loosening and tissue inflammation, on the bone regeneration. The optimal properties of the glass can be achieved by adjusting the glass compositions. The glass selected in this work was borosilicate glass. Its expansion coefficient is compatible with that of Al_2O_3 substrate. No crack was found in the coating (Fig. 4). The XRD pattern of the coating (Fig. 8) shows that there is no negative reaction between the glass and HA. Figure 8 is a typical HA diffraction pattern. Mechanical properties of HA-coated reticulated alumina are ~7–10.35 MPa for compressive strength and 5–8 MPa for tensile strength. Compared to previously reported porous materials, such as porous HA (1.3 MPa for compressive strength and 2.5 MPa for tensile strength) and coralline ceramic (5.8 MPa for compressive strength and 1.3 MPa for tensile strength) (6), a substantial increase in strength was obtained. These values are comparable to those of cancellous bone (2–12 MPa for compressive strength and 10–20 for tensile strength) (2). Some much stronger substrate materials, such as fiber reinforced composites, are excellent candidates for the HA coating using the developed approaches discussed in this article. The mechanical properties of the substrate can also be significantly improved by other ceramics routes. In addition, after bone ingrowth, the strength of the implant (bone composites) will be expected to further increase by a factor of 2–7 as previously demonstrated (7).

Thermal Deposition Method

Figure 9 shows the FTIR spectra of an unfired sample and samples fired at 500, 600, 700, and 900 °C. According to standard IR transmission spectra, peaks observed at 3573 and 631 cm^{-1} are assigned to OH stretching and librational modes. Peaks ~600 and 1100 cm^{-1} are due to the bending and stretching modes of PO bonds in the phosphate groups (8). These are characteristic peaks of HA. At a sintering temperature of 500 °C, PO bonds formed, but hydroxyl groups were not detected. Compared with the spectra of the unfired sample, most organic groups were burned out by this temperature. At 600 °C, all the characteristic lines of HA were recorded, but some organic residual could still be seen. At 700 °C and higher the peak positions match all those of standard HA, and the organic groups were not detectable.

Figure 10 shows the XRD spectra of HA sintered at different temperatures in the range of 600–900 °C. The results of the XRD are quite consistent with that of the FTIR. The crystalline phase started to form at 600 °C, and

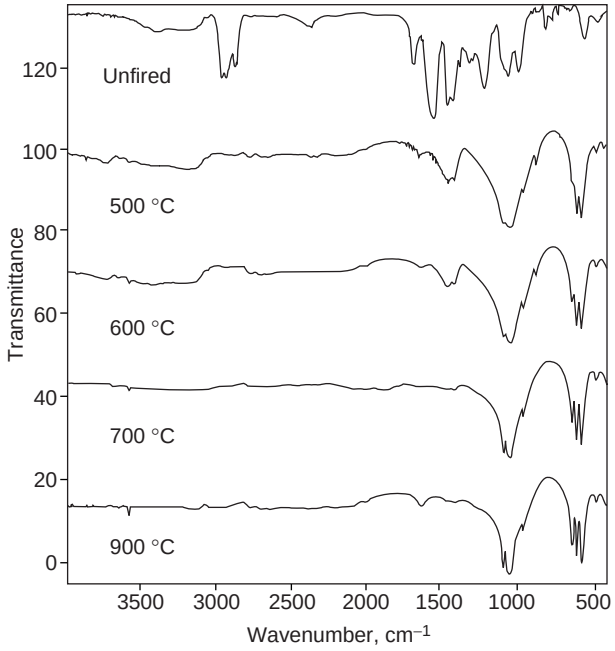


Figure 9. The FTIR spectra of HA-coated samples sintered for 4 h at the temperature indicated (thermal deposition).

all peaks were attributed to the HA phase. According to the Scherrer equation,

$$\Delta(2\theta) = K\lambda/D \tag{1}$$

where D is the crystallite dimension; K is the Scherer constant (here $K = 0.9$); is the X-ray wavelength in angstroms; $\Delta(2)$ is the true broadening of the diffraction peak at half-maximum intensity. The crystallite size is inversely proportional to the peak width. The broadening of peaks was evident at lower sintering temperatures, indicating the initial state of crystal formation. At higher sintering temperatures, the sharpening of peaks evidenced the growth of crystals. The peak shift could also be noted by comparing it with the standard XRD spectra of HA. At lower temperatures the shift was considerable, suggesting great lattice distortion.

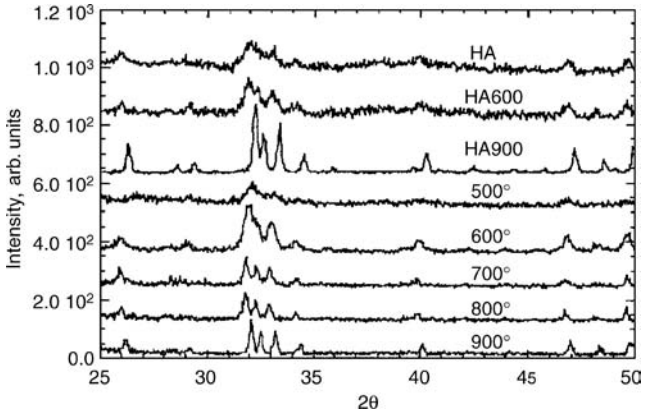


Figure 10. The XRD spectra of HA samples sintered at the temperatures indicated.

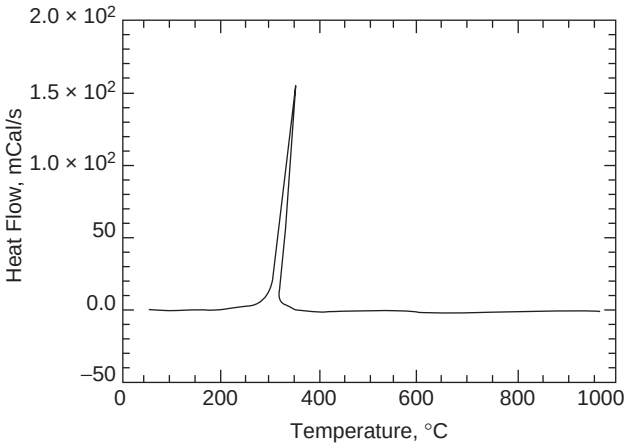


Figure 11. The DTA profile of the HA-coated sample showing a reaction near 300 °C.

Monitoring the sintering process and the evolution of chemical bonds is important in determining the bioactivity of the sintered products. The material with more lattice defects would be expected to be more reactive (2). This assumption will be experimentally verified later in *in vitro* tests. The high temperature and long sintering time will result in well-crystallized products. Therefore, to enhance the bioactivity, low temperature and short sintering time is preferred. It is critical to find an optimal sintering procedure so that the sintered HA is poorly crystallized but with no organic residuals. The DTA profile (Fig. 11) shows that burn-out of organic residuals occurs over the temperature range of 300–350 °C. In the current work, the samples were baked at 500 °C to burn out the organic groups. At this temperature, the structure of the sample is amorphous and most of the organic groups can be easily removed. This procedure will be helpful in eliminating residual carbon in the coating. Without this treatment, some organics could be incorporated in the final crystal lattice. It was found that the carbon disappeared at much lower temperatures than those samples treated in a rapid sintering process because most organic groups were not burned out at low temperatures. Therefore, a higher temperature is needed to remove them. However, the reactivity of HA is considerably reduced. It should be noted that the removal of the organic residue is not only related to the microstructure, but also to the macrostate of the samples. For example, for a thick and dense coating, a high temperature is needed to remove the residual carbon.

The bonding strength between the HA coating and the substrate was determined using the tape test. No peeling of the coating film was observed for all samples, indicating a strong bonding between the HA coating and the substrate. Figure 12 is the SEM micrographs showing the surfaces of the HA coating on dense alumina. As can be seen in this figure, the coating is fairly porous, which contributes to the bioactivity when immersed in SBF. Figure 13 is an SEM image of HA-coated reticulated alumina with significant open pores in the matrix. Figure 13b is the X-ray map recorded with Ca K_α lines for coated porous alumina. As can be seen in Fig. 13b, the distribution of calcium demonstrated that HA is uniformly coated on the skeleton of the

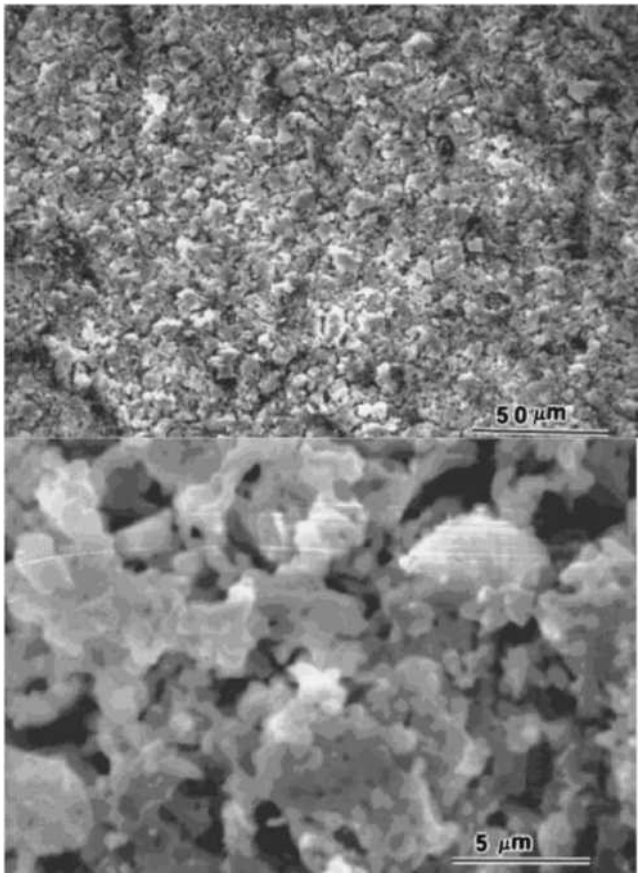


Figure 12. The SEM photographs showing the coated surface for a dense alumina substrate.

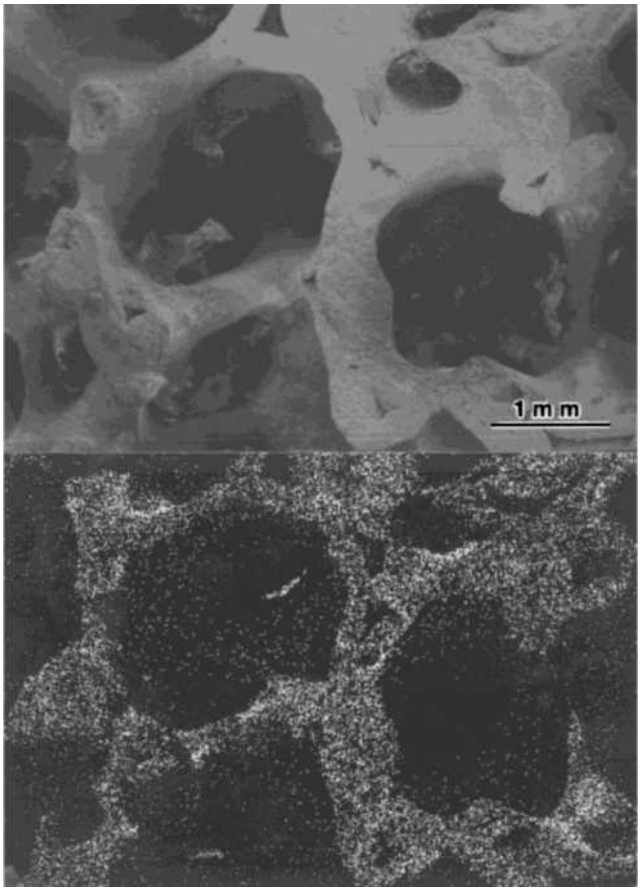


Figure 13. The SEM photographs showing (a) porous alumina substrate and (b) an X-ray map recorded with Ca K_α lines.

substrate. Figure 14 is the SEM micrograph showing the interface between the HA coating and the substrate. The coating thickness was ~ 1 μm, which can also be altered by a second or third coating.

Bioactivity Test

Due to bioactivity of HA, dissolution occurs after the sample is immersed in SBF. Consequently, some of the elements such as calcium in the solution are expected to change. The elemental-concentration changes of calcium in the SBF solution as a function of time are given in Figs. 15,16. As can be seen in Fig. 15, both HA and HA600 exhibit an immediate uptake of the Ca concentration. Initially, there is a high rate of ion uptake, suggesting the formation of a new phase on the HA surface in supersaturated solution. After 24 h, with the depletion of supersaturation, the reaction proceeds at a lower rate of uptake. For HA900, there is an induction time of 60 min prior to a detectable decrease in Ca concentration, and the initial rate of Ca uptake is much lower than those of HA and HA600. The SHA700 behaves similarly to HA and HA600 with a slow reaction rate, as can be seen in Fig. 16. However, the reaction behaviors of SHA800 and SHA900 significantly differ from the HA samples. During the first hour, an increase of Ca concentration was measured, indicating that dissolution of SHA800 and SHA900 has surpassed the new phase formation. Note that the rise in supersaturation for SHA900 is greater than that for SHA800. The ion uptake takes place after this initial dissolution. Another difference between HA and SHA series is that the latter took longer to reach the solid–solution equilibrium stage, clearly indicating a slower reaction rate in the HA series. These results suggest that the dissolution and precipitation rates are

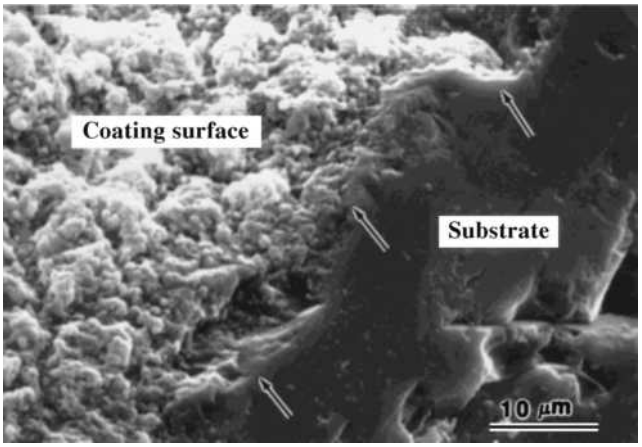


Figure 14. The SEM photograph showing the interface between the HA coating and the dense alumina substrate.

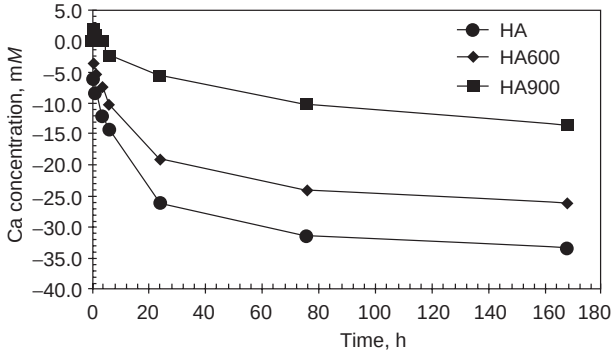


Figure 15. The Ca concentration in SBF versus immersion time for the HA group sintered at temperature indicated.

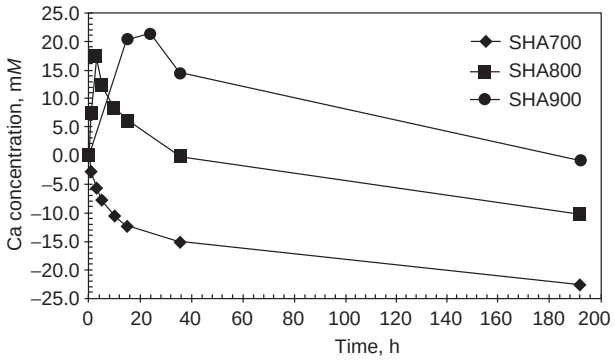


Figure 16. The Ca concentration in SBF versus immersion time for the SHA group sintered at temperature indicated.

critically dependent on the crystal structures developed in the HA samples.

Figure 17 represents the Ca concentration in the solution as a function of immersion time for samples Nos. 1, 2, and 3. All these samples are commercial HA heat treated at 700 °C for 30 min. Therefore, these samples are of the same structural crystallinity, but with different specific surface areas. Sample No. 3 has the highest SSA; Sample No. 1 has the lowest SSA; and Sample No. 2 is in the middle. They were tested at a ratio of 1 mg · mL^{-1} SBF. It is apparent in Fig. 17 that the rates of precipitation are highly dependent on the surface area. From these kinetic curves, the initial

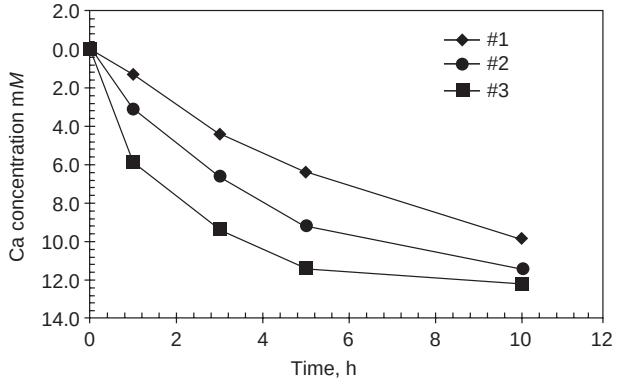


Figure 17. The Ca concentration in the solution as a function of immersion time for samples No. 1, No. 2, and No. 3.

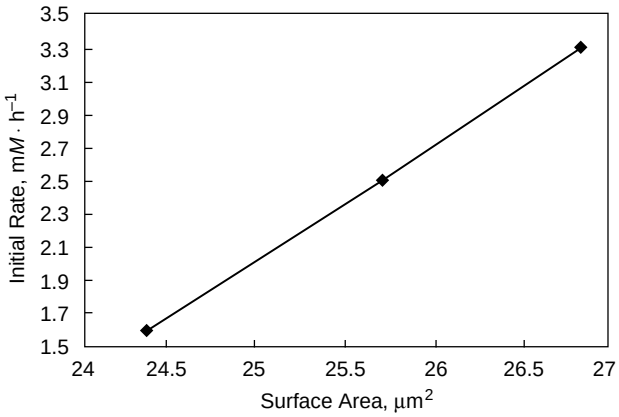


Figure 18. Initial reaction rate versus specific area for the samples showing in Fig. 17.

rate of precipitation, R_0, was determined by the slope of the first two data points. As shown in Fig. 18, there is a linear relationship between the initial precipitation rate and the surface area.

Figure 19 is a plot of Ca concentration versus immersion time for HA, HA600, SHA800, and SHA900. Samples of each group have been selected to have the same surface area. As can be seen, the initial rates of HAs and SHAs separate into two branches. The HA group exhibits an initial gradual decrease, while that of SHA group increases quite rapidly. However, as can also be seen in this figure, calcium concentration of SHA800 initially increases, but reaches a peak at 3 h, and thereafter decreases. In SHA900, although with a different rate, the calcium concentration always increases up to 9 h. Therefore, it can be concluded that the specific surface area is not the only factor that affects the reaction behavior of various HA samples. As discussed later, the degree of crystillinity in fact plays an even more important role in the reaction rates. The SHA700 sample behaves similarly to HA and HA600 with a slow reaction rate as can be seen in (Fig. 19). However, the reaction behaviors of SHA800 and SHA900 significantly differ from the HA samples. During the first hour, an increase of Ca concentration was measured indicating that dissolution of SHA800 and SHA900 has surpassed the new phase formation. It is noted that the rise in super-

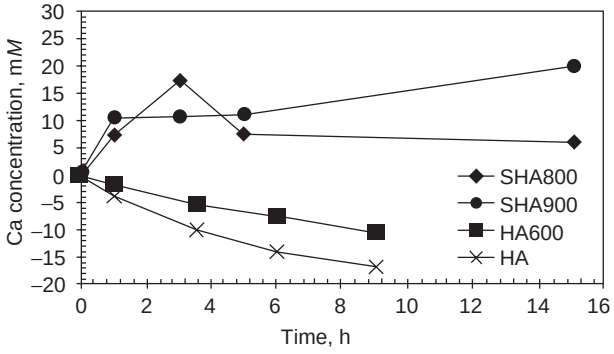


Figure 19. The Ca concentration versus immersion time for some of the typical HA and SHA samples.

Figure 20. The SEM photograph showing the surface morphology of the SHA700 immersed in SBF for 1 week.

saturation for SHA900 is greater than that for SHA800. The ion uptake takes place after this initial dissolution. Another difference between HA and SHA series is that the latter took longer time to reach the solid–solution equilibrium stage, clearly indicating a slower reaction rate in the HA series. These results suggest that the dissolution and precipitation rates are critically dependent on the crystal structures developed in the HA samples.

Figures 20 and 21 are surfaces of SHA700 and SHA900 coatings after immersion in SBF for 1 week and 9 weeks, respectively. As can be seen, the morphology of these two surfaces is quite different. The SHA700 coating is fully converted to flake shape with an average diameter of 1–10 μm. The surface of the flakes exhibits fine, needle-like structures within 1 week, which have been identified to be HCA by IR analysis. For SHA900 coating, after 9 weeks of immersion in SBF, a layer of precipitation has been observed under high magnification, which is shown to be amorphous or poorly crystallized new phase instead of HCA.

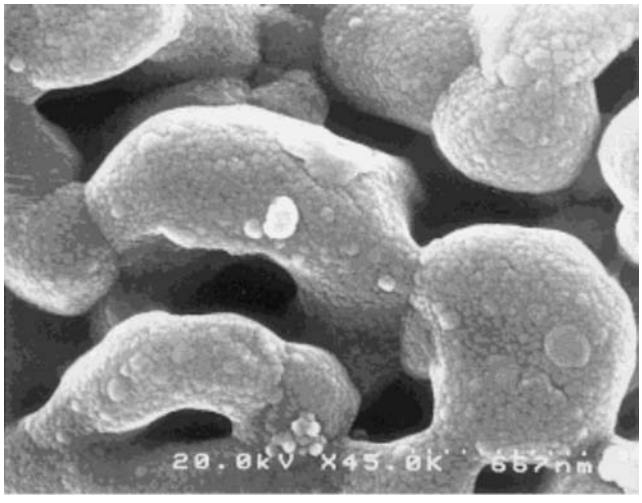


Figure 21. The SEM photograph showing the surface morphology of the SHA900 immersed in SBF for 9 weeks.

Figure 22 shows the FTIR spectra of HA and SHA samples after immersion in SBF at time periods up to 1 week. The absorption bands at 1460 cm^{-1} (high C=O region) and 872 cm^{-1} (low CO region) are characteristic features of HCA (8). As can be seen from spectrum of HA (Fig. 22a), these bands became significantly greater after 76 h of immersion indicating an increase in carbonate content. A gradual reduction of the splitting of the major PO_4^{3-} absorption bands (1100–1000 and 600–550 cm^{-1}) with immersion time is also observed, suggesting the formation of amorphous or fine, poorly crystallized new phases. For HA900, a broad band appears in the high energy C=O region (Fig. 22b). However, the low energy C=O band at 872 cm^{-1} is not recorded. At the same time, a gradual reduction of the splitting of the major PO_4^{3-} bands is observed, indicating again the formation of amorphous or fine, poorly crystallized new phases. The HCA phase cannot be identified from these weak bands, and it is likely that some intermediate phases other than HCA have formed. The HCA peaks have appeared in the spectra of SHA700 within 7 days (Fig. 22c). A time-dependent increase in the carbonate band intensities accompanied by a reduction of splitting of the major PO_4^{3-} bands is again recorded. Similar changes have occurred in the spectra of immersed SHA800 (Fig. 22d). However, no characteristic HCA peaks are recorded for SHA900 up to 3 weeks, only a broad band has appeared in the high energy C=O region (Fig. 22c). This trend seems to indicate that the reactivity of HA is considerably reduced at higher temperatures.

DISCUSSION

Structural Effects

The results indicate that *in vitro* behavior of the HA coatings is strongly affected by the structural characteristics induced by heat treatment. The SBF used in this work represents physiological ion concentration in human body, and it is supersaturated with respect to HA. In this chemical environment, HA is the most stable phase among all the calcium phosphate phases, thus the HCA formation is thermodynamically possible. However, only HA, HA600, and SHA700 have led to immediate Ca ions uptake. The HA900, SHA800, and SHA900 samples show a partial dissolution prior to precipitation. The difference in the dissolution ability of the HA samples is not the only factor in bioactivity. Figure 10 shows XRD spectra of HA sintered at different temperatures in the range of 600–900 °C. The structural evolution begins from an amorphous state in the commercial HA. Crystalline phase started to form at 600 °C, and all peaks were attributed to the HA phase. In addition, relative peak intensities are in agreement with the expected values for HA. Therefore, it can be decided that the structure consists primarily of crystalline HA, no additional peaks were observed to appear at any firing temperatures. However, the peak shift could be noted by comparing with the standard XRD spectra of HA. At lower temperatures, the shift was considerable suggesting great lattice distortion. The breadth of the peaks was used as an indicator of crystal dimension in the direction perpendicular to the diffracting plane hkl. The crystal size D is

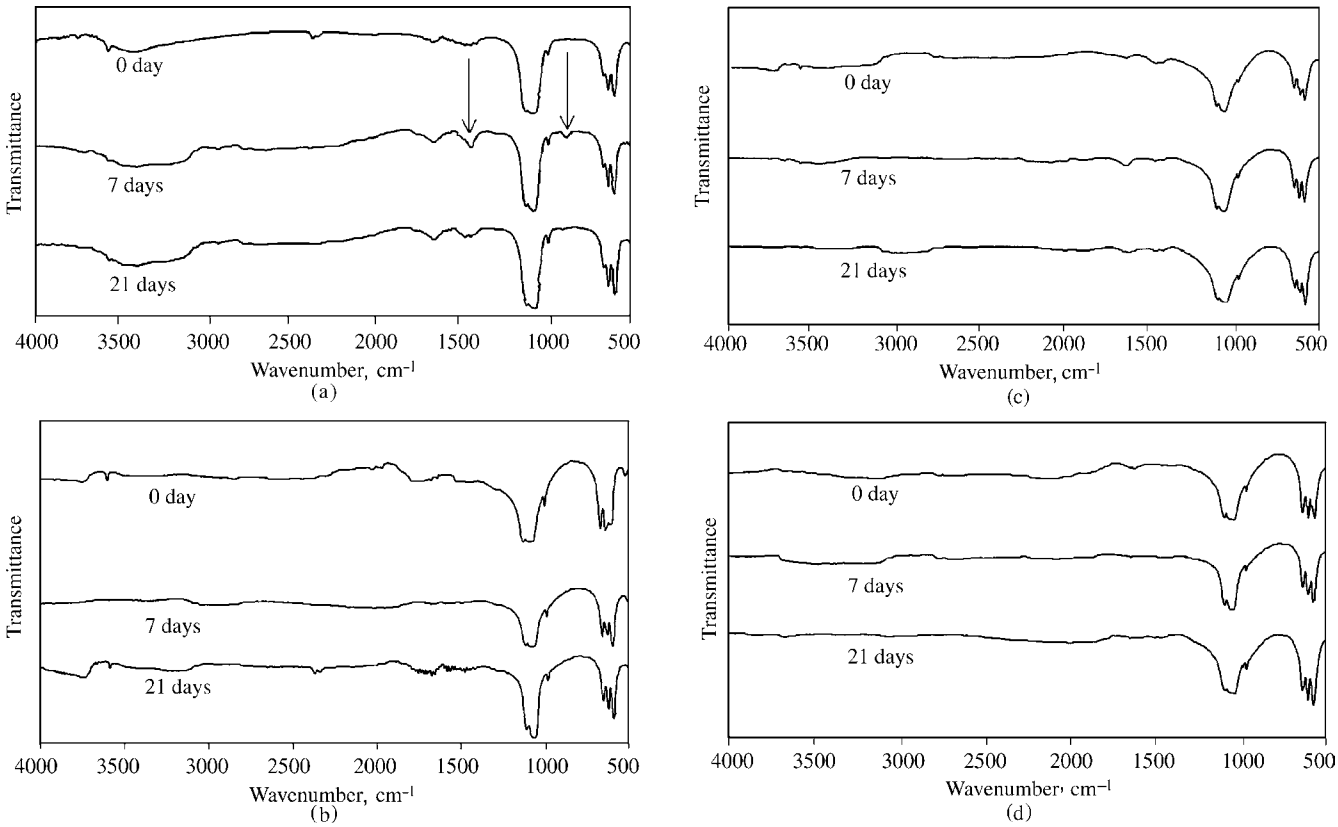


Figure 22. The FTIR spectra of HA and SHA samples after immersion in SBF for the time indicated.

inversely proportional to the peak breadth according to the Scherer equation. The contribution to the peak breadth from instrumental broadening was determined to be $\sim 0.12\,^{\circ}C$ (0.002 rad), independent of 2θ. This amount subtracted from the total experimental width is the value of true broadening, assuming the two contributions add linearly. The peak breadth (D002) is given as a function of temperature in (Fig. 23). It can be seen that the peak breadth decreases with sintering temperature, indicating that the crystal size increases with increasing sintering temperature, from 600 to 900 °C. On the basis of above analysis, the important difference with annealing temperature was the size of the individual crystals and the amount of crystal defects.

It is possible that the crystal growth rate is controlled by more than one of the elementary rate controlling mechanisms. The rate controlling process can change depending on particle size, solution concentration, and surface properties of the crystallites. The mechanisms of crystal growth are usually interpreted from measured reaction rates at different driving forces or from the activation energies of reactions. It is common practice to fit the data to an empirical rate law, which is represented by simple empirical kinetics (9):

$$R_g = k_g s \sigma^n \qquad (2)$$

where k_g is the rate constant for crystal growth, s is a function of the total number of available growth sites, is the degree of supersaturation, and n is the effective order of reaction. A broad empirical test for growth mechanism can be achieved from a logarithmic plot of Eq. 2. From the n value, the probable mechanism can be deduced. It is possible that the crystal growth rate is controlled by more than one of the elementary rate controlling mechanisms listed above. Under these circumstances, the rate-limiting steps are dependent on the jump frequency of lattice ions: (*1*) through the solution for mass transport control; (*2*) to the crystal surface for adsorption control, or (*3*) along the crystal surface or into a crystal lattice kink site for spiral and polynuclear control. The rate controlling process can change depending on particle size, solution concentration, and surface properties of the crystallites. A broad empirical

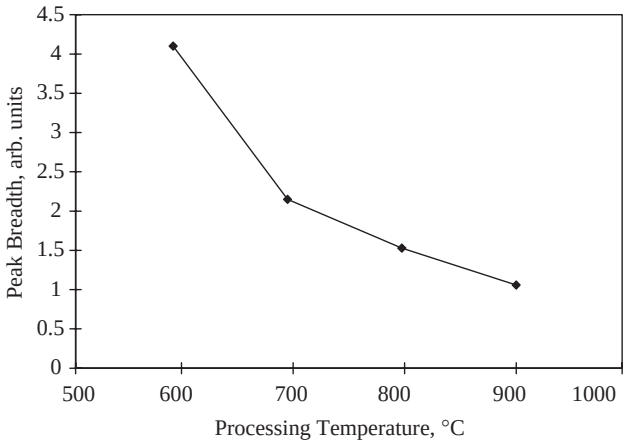


Figure 23. The X-ray diffraction peak breadth versus processing temperature for the HA synthesized in this study.

test for growth mechanism can be achieved by plotting the data according to Eq. 2. An effective order reaction in the range $0 < n < 1.2$, $n \sim 2$, or $n > 2.5$ indicates that the rate controlling process is one of adsorption and /or mass transport, surface spiral, or polynucleation, respectively (9). Experimentally, it is found that the growth rates of the calcium phosphates are insensitive to changes in fluid dynamics indicating surface controlling mechanisms rather than mass transport of ions to the crystal surfaces. A dynamic fluid may effect the growth rate, but not in a pronounced fashion. However, in our experiment, we found the growth to be insensitive. But we have no comparison of growth rates in both static and dynamic fluids.

Temperature Dependence of Activation Energy

Activation energies, obtained from experiments at different temperatures, can be used to differentiate between volume diffusion and surface controlled processes. The activation energy for volume diffusion, reflecting the temperature dependence of the diffusion coefficient, usually lies between 16 and 20 $kJ \cdot mol^{-1}$, while for a surface reaction the value may be in excess of 35 $kJ \cdot mol^{-1}$. If a reaction has activation energy of < 20 $kJ \cdot mol^{-1}$, it is safe to assume that it is overwhelmingly controlled by volume diffusion. However, if the activation energy is higher than 35 $kJ \cdot mol^{-1}$, it is quite certain that an adsorption process predominates. In all other cases, both adsorption and volume diffusion mechanisms may participate for a first-order reaction (10).

Figure 17 represents the Ca concentration in solutions as a function of immersion time at different surface areas. These samples were the same kind of powders to ensure that they have the same crystal structure and surface morphology; while the ratio of surface area to volume of SBF was different. It is apparent in Fig. 17 that the rates of precipitation were highly dependent on the surface area. Based on the empirical kinetics in Eq. 2, to build a relationship between the reaction rate R_g and surface area s, the degree of supersaturation σ should be kept at a constant value. The corresponding reaction rates were calculated by a simple fitting procedure from the above kinetic plots. As shown in Fig. 18, there was a linear relationship between the precipitation rates and the total surface area, which is in agreement with the above empirical kinetics equation. This result also showed that crystallization occurred only on the added seed materials without any secondary nucleation or spontaneous precipitation. Furthermore, the advantage of porous bioceramics over dense bioceramics was proved by this relationship.

The initial precipitation rate was not used here because of the following considerations. First, the empirical fitting procedure used to calculate R_0 is greatly affected by the slower rates occurring after the initial fast stage of the precipitation process. Thus, the fitting data could not represent the true initial rate. Second, initial rate was a complicated factor. Rapid adjustment of surface composition usually happened when the solids were introduced into the growth or dissolution media. In the case of HA, initial uptaking surges were observed, which might be attributed to calcium ion adsorption. Therefore, considerable uncertainties can arise if too much emphasis is placed upon initial rates of reaction.

Another point needed to be noted was that in this test, the different surface areas were not originated from the distribution of particle sizes, considering that different particle sizes might bring in the factor of surface morphology, which has great influence on the reaction rate. The effect of particle size would be demonstrated later. In the current method, the same powders were used, so that the factor of morphology was eliminated and a linear relationship was obtained.

The particles of different sizes behaved differently under the same surface area to volume (SA/V) test conditions. When comparing the 40–100-mesh and < 200-mesh particles at SA/V of 0.02 $m^2 \cdot mL^{-1}$, it is apparent that the Ca adsorption rate is slower for the smaller particles. This may be attributed to physical differences such as the radius of curvature and surface roughness.

Figure 19 is a plot of Ca concentration verses immersion time for HA, HA600, and SHA800, SHA900. Samples of each group were tested under the same SA/V ratio. As can be seen, the initial rate of HA was greater than that of HA600; the behavior of SHA800 differed from the one of SHA900. Therefore, it is concluded that the specific surface area is not the only reason that affects the reaction behavior of various HA powders, the degree of crystallinity also plays an important role in their reaction rates.

Chemical reactions, specifically in this case, the process of nucleation and crystal growth from solution, is described as an activated process with temperature, which is represented by the following relationship:

$$\text{rate} \propto \exp\left(-\frac{E_a}{kT}\right) \tag{3}$$

where E_a is the activation energy, so that reaction rate increases exponentially with temperature increase. The reaction rate constant K is related to temperature by an Arrhenius equation:

$$K = K_0 \exp\left(-\frac{E_a}{kT}\right) \tag{4}$$

By keeping σ at a constant value, plot $\ln R$ versus $1/T$, the slope of the curve will be E_a/k, and consequently E_a can be calculated.

According to the procedures described above, the activation energy for HA, HA600, HA900, and SHA700 was calculated. The parameter σ was selected at $\Delta Ca = -8mM$ for all the reaction temperatures. The computed activation energy was listed in Table 3. The above results showed that the activation energy increased with the sintering temperature for HA powders. The activation energy of synthesized HA700 was much higher than those of HA and HA600.

In Vitro Biochemistry Behavior of Hydroxyapatite

The formation of biological apatite on the surface of implanted synthetic calcium phosphate ceramics goes

Table 3. Activation Energies for the Samples Indicated

Samples	Activation energy, $kJ \cdot mol^{-1}$
HA	66.3
HA600	80.3
HA900	172.7
SHA700	130.4

through a sequence of chemical reactions. It has been shown that the reaction rate *in vitro* appears to correlate with the rate of apatite mineral formation *in vivo*.

Therefore, the laboratory observations can be projected to the *in vivo* situation. The *in vitro* behavior of bioceramics is determined by its stability at ambient and body temperatures. Many factors have significant influence on their stability, including the pH and supersaturation of the solution, crystallinity, structure defects, and porosity of the material (11,12). Driessens (13) showed that, among the phases composed of calcium and phosphate, hydroxyapatite is the most stable at room temperature when in contact with SBF, which was used to represent the ionic concentrations of plasma. Generally, SBF will initiate a partial dissolution of the HA material causing the release of Ca^{2+}, HPO_4^{2-}, and PO_4^{3-}, and increasing the supersaturation of the microenvironment with respect to HA phase that is stable in this environment. Following this initial dissolution is the reprecipitation. Carbonate ions, together with other electrolytes, which are from the biological fluids, become incorporated in the new apatite microcrystals forming on the surfaces of the HA.

Since any clinical use of calcium phosphate bioceramics involves contact with water, it is important to understand the stability of HA in the presence of water at ambient temperatures. As Driessens showed (13), there were only two classes of calcium phosphate materials that were stable at room temperature when in contact with aqueous solution. Temperature, ionic strength, and pH are major parameters influencing the stability of calcium phosphate. In the body, temperature and ion strength are constant, therefore, the pH value at the local tissue determines which form of calcium phosphate is the most stable. At a pH < 4.2, the component $CaHPO_4{\cdot}2H_2O$ was the most stable, while at higher pH (> 4.2), HA was the stable phase. However, HA does not form at the first place. Other mineral phases such as dicalcium phosphate dihydrate (DCPD), octacalcium phosphate (OCP), and amorphous tricalcium phosphate (TCP) form as precursor phases that transform to HA.

Therefore, in this *in vitro* test, at biological pH value, only HA or its precursor phase can be found in contact with SBF. It is believed that synthetic HA ceramic surfaces can be transformed to biological apatite through a set of reactions including dissolution, precipitation, and ion exchange. Following the introduction of HA to SBF, a partial dissolution of the surface is initiated causing the release of Ca^{2+}, HPO_4^{2-}, and PO_4^{3-}, which increases the supersaturation of the microenvironment with respect to the stable (HA) phase. Carbonated apatite can form using the calcium and phosphate ions released from partially dissolving ceramic HA and from the biological fluids that contain other electrolytes, such as CO_3^{2-} and Mg^{2+}. These become incorporated in the new CO_3-apatite microcrystals forming on the surfaces of ceramic HA crystals. The *in vitro* reactivity of HA is governed by a number of factors, which can be considered from the two aspects: *in vitro* environment and properties of HA material.

CONCLUSIONS

In order to produce highly strengthened porous bioactive materials for bone substitutes, suspension method and thermal deposition method, were employed to coat the inner-pore surfaces of a porous ceramic substrate. A thin layer of HA has been uniformly coated onto inner-pore surfaces of reticulated alumina substrates. The *in vitro* bioactivity of HA coatings was found to be strongly affected by structure characteristics, which are a combination of crystallinity and specific surface area. The bioactivity is reduced at a higher degree of crystallinity, which is likely related to the higher driving force for the formation of a new phase, and the reaction rate was proportional to the surface area. The surface morphology and HA treating temperature also have a direct affect on the reaction rates of the HA coatings. The calcium absorption rate is slower for smaller particles; this could be attributed to physical differences including radius of curvature and surface roughness. The activation energy increased with the heat-treatment temperature for HA powders.

BIBLIOGRAPHY

Cited References

1. Shinzato S, et al. Bioactive bone cement: Effect of silane treatment on mechanical properties and osteoconductivity. J Biomed Mater Res 2001;55(3):277–284.
2. Hench LL. Introduction to Bioceramics. Singapore: World Scientific; 1993. p 139–180.
3. Barth E, Hero H. Bioactive glass ceramic on titanium substrate: the effect of molybdenum as an intermediate bond coating. Biomaterials 1986;7(4):273–276.
4. Kasuga T, et al. Bioactive calcium phosphate invert glass-ceramic coating on beta-type Ti-29Nb-13Ta-4.6Zr alloy. Biomaterials 2003;24(2):283–290.
5. Livingston T, Ducheyne P, Garino J. In vivo evaluation of a bioactive scaffold for bone tissue engineering. J Biomed Mater Res 2002;62(1):1–13.
6. Roy DM, Linnehan SK. Hydroxyapatite formed from coral skeletal carbonate by hydrothermal exchange. Nature (London) 1974;247(438):220–222.
7. Holmes R, et al. A coralline hydroxyapatite bone graft substitute. Preliminary report. Clin Orthop 1984;188:252–262.
8. Radin SR, Ducheyne P. The effect of calcium phosphate ceramic composition and structure on in vitro behavior. II. Precipitation. J Biomed Mater Res 1993;27(1):35–45.
9. Nielsen AE. Electrolyte Crystal Growth Mechanisms. J Crystal Growth 1984;67: 289–310.
10. Gengwei J. Development of Bioactive Materials using Reticulated Ceramics for Bone Substitue. Ph. D. dissertation, University of Cincinnati; 2000. p 118.
11. Margolis HC, Moreno EC. Kinetics of hydroxyapatite dissolution in acetic, lactic, and phosphoric acid solutions. Calcif Tissue Int 1992;50(2):137–143.
12. Christoffersen J, Christoffersen MR. Kinetics of Dissolution of Calcium Hydroxyapatite. 5. The Acidity Constant for the Hydrogen Phosphate Surface Complex. J Crystal Growth 1982;57: 21–26.
13. Driessens FCM. Formation and Stability of Calcium Phosphates in Relation to the Phase Composition of the Mineral in Calcified Tissues. In: de Groot K, editor. Bioceramics of Calcium Phosphate. Boca Raton, (FL): CRC Press; 1983; p 1–31.

See also BIOMATERIALS, CORROSION AND WEAR OF; NUCLEAR MAGNETIC RESONANCE SPECTROSCOPY; POROUS MATERIALS FOR BIOLOGICAL APPLICATIONS.

BIOMATERIALS: TISSUE-ENGINEERING AND SCAFFOLDS

GILSON KHANG
Chonbuk National University
SANG JIN LEE
MOON SUK KIM
HAI BANG LEE
Korea Research Institutes of Chemical Technology

INTRODUCTION

Tissue engineering offers an alternative to whole organ and tissue transplantation for diseased, failed, or abnormally functioning organs. Millions suffer from end-stage organ failure or tissue loss annually. In the United State alone, at least 8 million surgical operations are carried out each year at a total national healthcare cost exceeding $400 billion annually (1–4). Approximately 500,000 coronary artery bypass surgeries are conducted in the United States annually (5). Autologous and allogenic natural tissue, such as the saphenous vein or the internal mammary artery, is generally used for coronary artery replacement. The results have been favorable for these procedures with patency rates generally ranging from 50–70%. Failures are caused by intimal thickening due largely to adaptation of the vessel in response to increased pressure and wall shear stress, compression, inadequate graft diameter, and disjunction at the anastomosis. Also, successful treatment has been limited by the poor performance of the synthetic materials used, such as polyethyleneterephthalate (PET, Dacron) and expanded polytetrafluoroethylene (ePTFE, Gore-Tex), which are used for tissue replacement due to plaguing problems (6). For example, in cases of tumor resection in the head, neck, and upper and lower extremities, as well as in cases of trauma and congenital abnormalities, there are often outline defects due to the loss of soft tissue, this tissue is largely composed of subcutaneous adipose tissue (7). The defects lead to abnormal cosmesis, affect the emotional comfort of patients, and may impair function. A surgeon would prefer to use an autologous adipose tissue to sculpt contour deformities. Because mature adipose tissue does not transplant effectively, numerous natural, synthetic, and hybrid materials have been used to act as adipose surrogates. Despite improved patient outcomes, the use of many of these materials results in severe problems, such as unpredictable outcomes, fibrous capsule contraction, allergic reactions, suboptimum mechanical properties, distortion, migration, and long-term resorption.

To offset the short supply of donor organs as well as the problems caused by the poor biocompatibility of the biomaterials used, a new hybridized method of "tissue engineering", which combines both cells and biomaterials has been introduced (8). To reconstruct new tissue by tissue engineering, a triad of components are requried: (*1*) harvested and dissociated *cells* from the donor tissue including nerve, liver, pancreas, cartilage, and bone as well as embryonic stem, adult stem, or precursor cell; (*2*) *scaffolds* made of biomaterials on which cells are attached and cultured, then implanted at the desired site of the functioning tissue; (*3*) *growth factors* that promote and/or prevent cell adhesion, proliferation, migration, and differentiation by up-regulating or down-regulating the synthesis of protein, growth factors, and receptors (see Fig. 1). In a typical application for cartilage regeneration, donor cartilage is harvested from the patient and dissociated into individual chondrocyte cells using enzymes as collagenase, and then mass cultured *in vitro*. The chondrocyte cells are then seeded onto a porous and synthetic biodegradable scaffold. This cell–polymer structure is massively cultured in a bioreactor. The abnormal tissue is removed and the cell–polymer structure is then implanted in the patient. Finally, the synthetic biodegradable scaffold resorbs into the body and the chondrocyte cell produces collagen and glycosaminoglycan as its own natural extracellular matrix (ECM), which results in regenerated cartilage. This approach can theoretically be applied to the manufacture of almost all organs and tissues except for organs such as the brain (3).

In this section, a review is given of the biomaterials and procedures used in the development of tissue-engineered scaffolds, including: (*1*) natural and synthetic biomaterials, (*2*) natural–synthetic hybrid scaffolds, (*3*) the fabrication methods and techniques for scaffolds, (*4*) the required physicochemical properties for scaffolds, and (5) cytokine-released scaffolds.

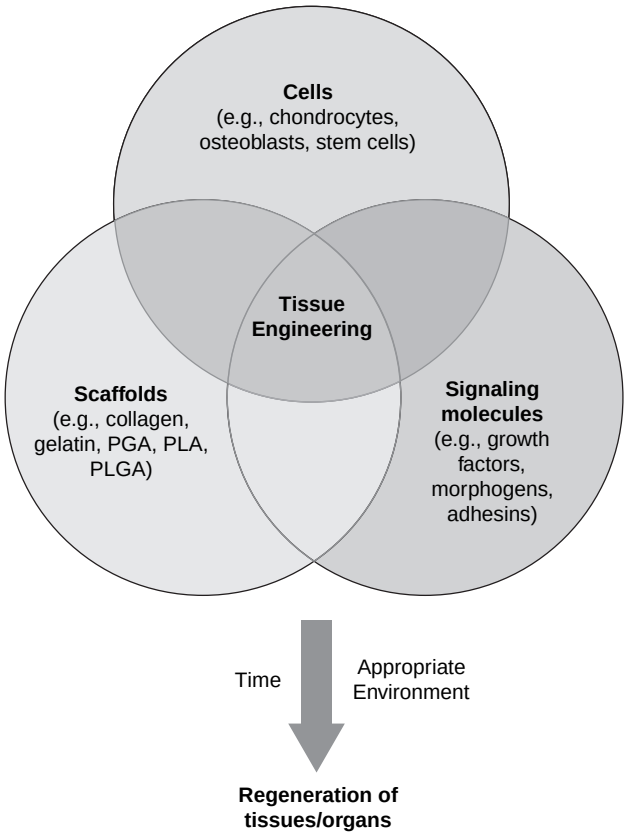


Figure 1. Tissue engineering triad. The combination of three key elements, cells, biomaterials, and signaling molecules, results in regenerated tissue-engineered neo-organs.

BIOMATERIALS FOR TISSUE ENGINEERING

The Importance of Scaffold Matrices in Tissue Engineering

Scaffolds play a very critical role in tissue engineering. Scaffolds direct the growth (*1*) of cells seeded within the porous structure of the scaffold, or (*2*) of cells migrating from surrounding tissue. Most mammalian cell types are anchorage dependent; the cells die if an adhesion substrate is not provided. Scaffold matrices can be used to achieve cell delivery with high loading and efficiency to specific sites. Therefore, the scaffold must provide a suitable substrate for cell attachment, cell proliferation, differentiated function, and cell migration. The prerequisite physicochemical properties of scaffolds are (*1*) to support and deliver the cells; (*2*) to induce, differentiate, and promote conduit tissue growth; (*3*) to target the cell-adhesion substrate, (*4*) to stimulate cellular response; (*5*) to create a wound healing barrier; (*6*) to be biocompatible and biodegradable; (*7*) to have relatively easy processability and malleability into the desired shapes; (*8*) to be highly porous with large surface–volume; (*9*) to have mechanical strength and dimensional stability; and (*10*) to have sterilizability (9–16). Generally, three-dimensional (3D) porous scaffolds can be fabricated from natural and synthetic polymers (Fig. 2 shows these chemical structures), ceramics, metal, and in a very few cases, composite biomaterials and cytokine-releasing materials.

Natural Polymers

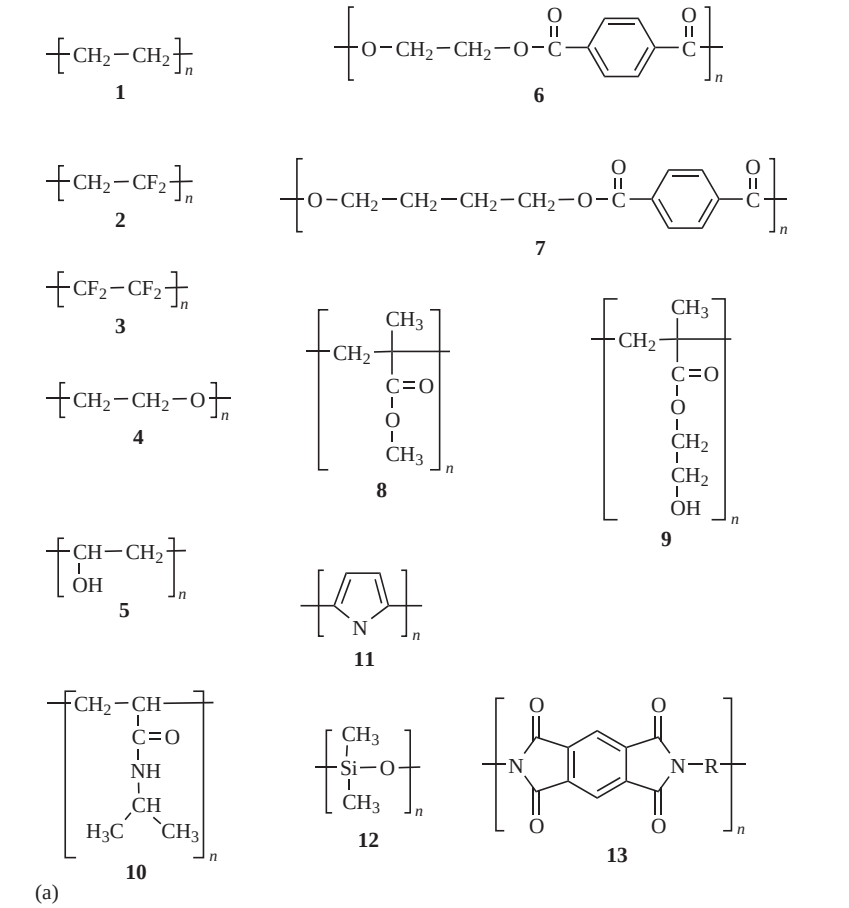
Many naturally occurring scaffolds can be used for tissue engineering purposes. One such example is the ECM, which is composed of very complex biomaterials and controls cell function. For the ECM used in tissue engineering, natural and synthetic scaffolds are designed to mimic specific function. The natural polymers used are alginate, proteins, collagens (gelatin), fibrins, albumin, gluten, elastin, fibroin, hyarulonic acid, cellulose, starch, chitosan (chitin), sclerolucan, elsinan, pectin (pectinic acid), galactan, curdlan, gellan, levan, emulsan, dextran, pullulan, heparin, silk, chondroitin 6-sulfate, polyhydroxyalkanoates, and others. Much of the interest in these natural polymers comes from their biocompatibility, relatively abundance and commercial availability, and ease of processing (17).

Alginate. Alginate (from seaweed) is composed of two repeating monosaccharides: L-guluronic acid and D-mannuronic acid. Repeating strands of these monomers form linear, water-soluble polysaccharides. Gelation occurs by interaction of divalent cations (e.g., Ca^{2+}, Mg^{2+}) with blocks of guluronic acid from different polysaccharide chains (as shown in Fig. 3). From this gelation property, the encapsulation of calcium alginate beads impregnated with various pharmaceutics, cytokines, or cultured cells, has been extensively investigated. Varying the preparation conditions of the gelation can control structure and physicochemical properties. Calcium alginate scaffolds do not degrade by hydrolytic reaction, whereas they can be degraded by a chelating agent such as ethyleneaminetetraaceticacid (EDTA) or by an enzyme. Also, the diffusion of calcium ions from an alginate gel can cause dissociation between alginate chains, which results in a decrease of mechanical strength over time. One of the disadvantages of an alginate matrix is a potential immune response and the lack of complete degradation, since alginate is produced in the human body (10). For these reasons, the chemical modification and incorporation of biological peptides, such as Arg-Gly-Asp cell adhesion peptides, have been used to improve the functionality and flexibility of natural scaffolds and their potential application (18).

Many researchers have studied the encapsulation of chondrocytes. Growth plate chondrocytes, fetal chondrocytes, and mesenchymal stem cells derived from bone marrow have been encapsulated in alginate (19). In each system, the chondrocytes demonstrated a differentiated phenotype, producing an ECM and retaining the cell morphology of typical chondrocytes. In addition, novel hybrid composites, such as alginate/agarose (a thermosensitive polysaccharide), alginate/fibrin, alginate/collagen and alginate/hyaruronic acid, and different gelling agents (water, sucrose, sodium chloride, and calcium sulfate) were investigated to optimize the advantages of each component material for tissue engineered cartilage (20–22). It was found that this hybrid material provides a reason why the microenvironments of composite materials affect chondrogenesis.

Collagen. At least 22 types of collagen exist in the human body. Among these, collagen types I, II, and III are the most abundant and ubiquitous. Conformation of the collagen chain consists of triple helices that are packed or processed into microfibrils. Molecularly, the three repeating amino acid sequences, such as glycine, proline, and hydroxyproline, form protein chains resulting in the formation of a triple helix arrangement. Type I collagen is the most abundant and is the major constituent of bone, skin, ligament, and tendon, whereas type II collagen is the collagen in cartilage. Collagen can promote cell adhesion as demonstrated by the Asp-Gly-Glu-Ala peptide in type I collagen, which functions as a cell-binding domain. Due to the abundance and ready accessibility of these tissues, they have been used frequently in the preparation of collagen (23).

The purified collagen materials obtained from either molecular or fibrillar technology are subjected to additional processing to fabricate the materials into useful scaffold types for specific tissue-engineered organs. Collagen can be processed into several types such as membrane (film and sheet), porous (sponge, felt, and fiber), gel, solution, filamentous, tubular (membrane and sponge), and composite matrix for the application of tissue repair, patches, bone and cartilage repair, nerve regeneration, and vascular and skin repair with/without cells (24). The Physicochemical properties of collagen can be improved by the addition of a variety of homogeneous and heterogeneous composites. Homogeneous composites can be formed between ions, peptides, proteins, and polysaccharides in a collagen matrix by means of ionic and covalent bonding, entrapment, entanglement, and coprecipitation. Heterogeneous composites, such as collagen–synthetic polymers, collagen–biological polymers, and collagen–ceramic hybrid



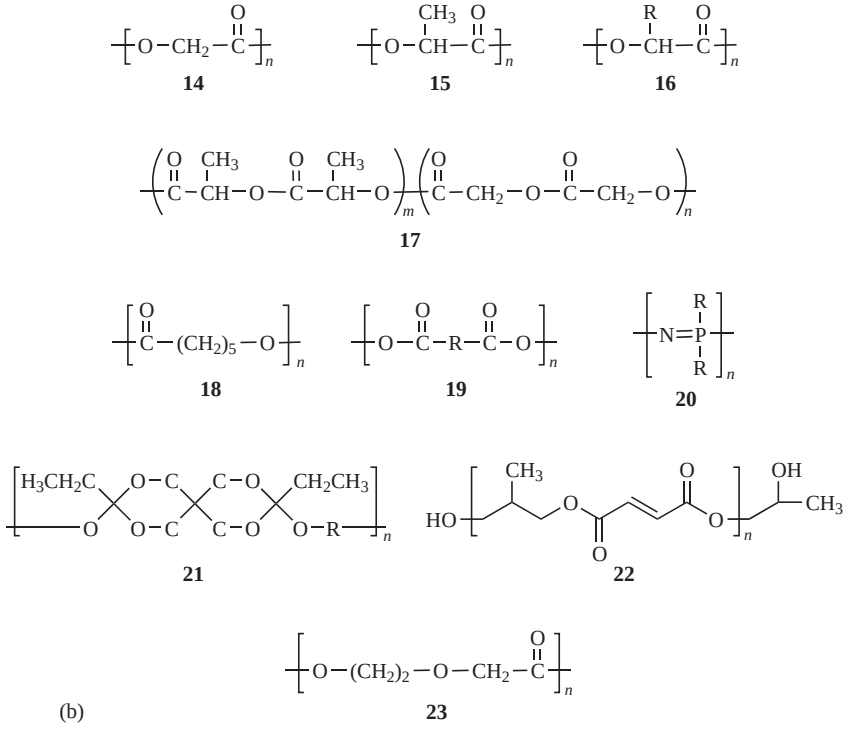
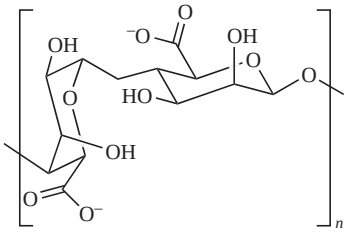
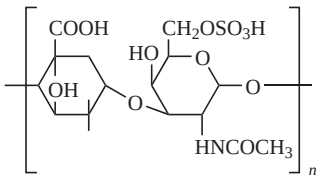


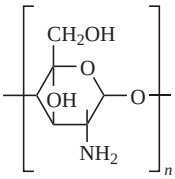
Figure 2. Chemical structures of some commonly used biodegradable and nondegradable polymers in tissue engineering. (a) Synthetic nondegradable polymers: (**1**). polyethylene, (**2**). poly(vinylidene fluoride), (**3**). polytetrafluoroethylene, (**4**). poly-(ethylene oxide), (**5**). poly(vinyl alcohol), (**6**). poly-(ethyleneterephthalate), (**7**). poly(butyleneterethphalate), (**8**). poly(methylmethacrylate), (**9**). poly- (hydroxymethylmetacrylate), (**10**). poly(*N*-isopropylacrylamide), (**11**). polypyrrole, (**12**). poly(dimethyl siloxane), and (**13**). polyimides. (b) Synthetic biodegradable polymers: (**14**). poly(glycolic acid), (**15**). poly(lactic acid), (**16**). poly(hydroxyalkanoate), (**17**). poly(lactide-*co*-glycolide), (**18**). poly(ε-caprolactone), (**19**). polyanhydride, (**20**). polyphsphazene, (**21**). poly(orthoester), (**22**). poly-(propylene fumarate), and (**23**). poly(dioxanone). (c) Natural polymers: (**24**). alginate, (**25**). chondroitin-6-sulfate, (**26**). chitosan, (**27**). hyarunonan, (**28**). collagen, (**29**). polylysine, (**30**). dextran, and (**31**). heparin. (d) PEO-based hydrogels: (**32**). Pluronic, (**33**). Pluronic R, (**34**). Tetronic, and (**35**). Tetronic R.



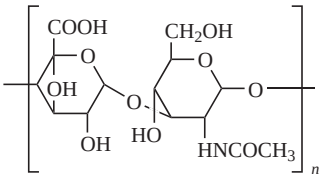
24



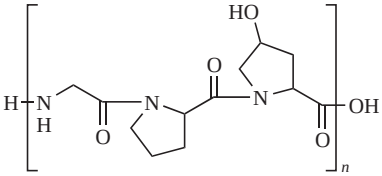
25



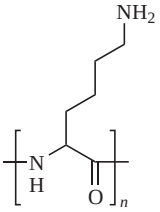
26



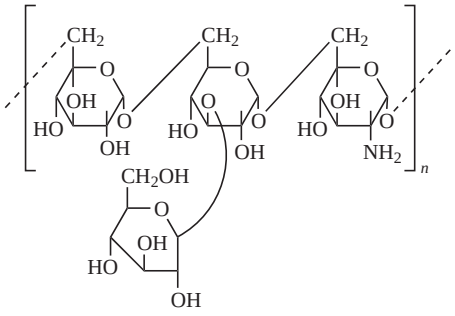
27



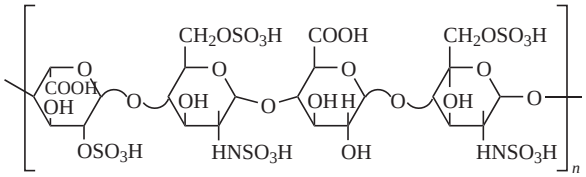
28



29

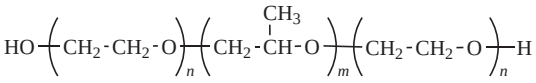


30

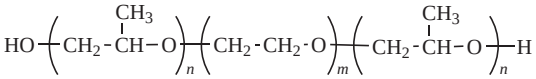


31

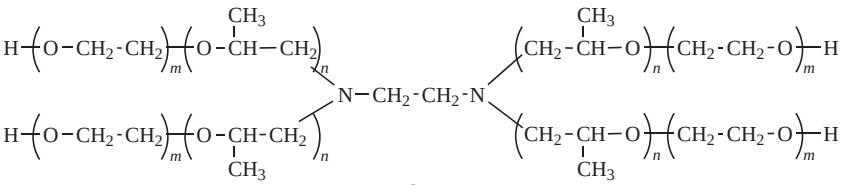
(c)



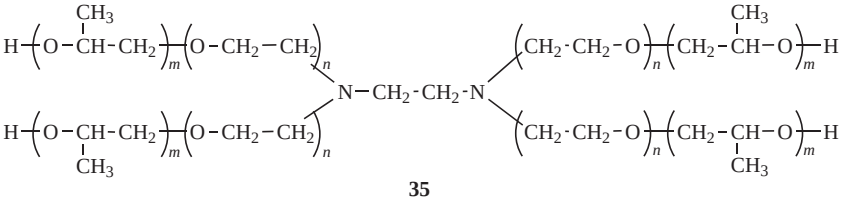
32



33



34



35

(d)

Figure 2. (*continued*)

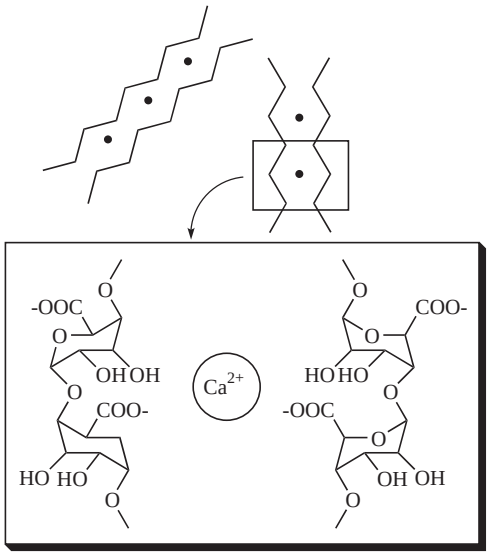


Figure 3. Schematic representation of the guluronate junction zone in alginate; eggbox model. The circles represent calcium ions.

polymers (collagen–nano-hydroxyapatite and collagen–calcium phosphate) have been investigated for use in tissue-engineered products (10).

Fibrin. Fibrin plays a major role during wound healing as a hemostatic barrier to prevent bleeding and to support a natural scaffold for fibroblasts. Actual polymerization is triggered by the conversion of fibrinogen to fibrin monomer by thrombin, and gelation occurs within 30–60 s. One advantage of using fibrin in this manner is its ability to completely fill the defect by gelling in situ. Fibrin sealant composed of fibrinogen and thrombin in addition to antifibronolytic agents has been used already in such surgical applications as sealing lung tears, cerebral spinal fluid leaks, and bleeding ulcers, because of its natural role in wound healing. Fibrin sealant might be made from autologous blood or from recombinant proteins (22). Fibrin gels can degrade either through hydrolytic or proteolytic means. Fibrinogen is commercially available from several manufacturers, so the cost of the fabrication of fibrin gels is relatively low. Recently, much work has been done to develop fibrin as a potential tissue-engineered scaffold matrix, especially for cartilage, which is formed from a fibrin/chondrocyte construct. Biochemical and mechanical analysis has demonstrated its cartilage-like properties. In neural tissue engineering, fibrin modified the incorporation of bioactive peptide in fibrin gels (25). Also, fibrin/hydroxyapatite hybrid composites have been investigated to optimize the mechanical strength of tissue-engineered subchondral bone substitutes.

Hyaluronan. Hyaluronic acid, a natural glycosaminoglycans polymer, can be found in abundance within cartilaginous ECM. It has some disadvantages in its natural form, such as high water solubility, fast resorption, and fast tissue clearance times, which are not conducive to biomaterials. To overcome these undesirable characteristics, chemical modifications were made to increase biocompatibility, tailor the degradation rate, control water solubility, and to fit the mechanical property. To increase hydrophobicity, esterification was carried out to increase the hydrocarbon content of the added alcohol, which resulted in tailored degradation rates since hydrophobicity directly influences hydration and the deesterification reaction (10). Another approach, the condensation reaction between the carboxylic group of unmodified hyaluronan molecules with the hydroxyl group of other hyalunonic acid molecules, was used to fabricate the sponge form. Then, bone marrow-derived mesenchymal progenitor cells were seeded to induce chondrogenesis and osteogenesis on this scaffold. Results from animal studies indicate that modified hyaluronic acid can successfully support mesenchymal stem cell proliferation and differentiation for osteochondral application (15). Also, a sulfate reaction on a hyaluronan gel created a variety of sulfate derivatives, ranging from one-to-four sulfate groups per disaccharide subunit. A crosslinking network hydrogel can be formed by using diamines from individual hyaluronic acid chains. Chondrocytes seeded on sulfated hyaluronic acid hydrogels appear to have good cell compatibility with the tissue-engineered cartilage. The benzyl ester hyaluronan products HYAFF-11 and LaserSkin (Fidia Advanced Biopolymers, FAB, Abano Terme, Italy) have been introduced to engineer skin bilayers *in vitro* (26).

Chitosan. Chitosan, a polysaccharide derived from chitin, is composed of a simple glucosamine monomer and has physicochemical properties similar to many glycosaminoglycans. Chitosan is relatively biocompatible and biodegradable; it does not evoke a strong immune response. It is relatively cheap due to its abundance and good reactivity with diverse methods of chemical processing. Chitin is typically extracted from arthropod shells by means of acid–alkali treatment to hydrolyze acetamido groups from the *N*-acetylglucosamine resulting in the production chitosan. It has a molecular weight of 800,000–1,500,000 $g \cdot mol^{-1}$ and dissolves easier than the native chitin polymer (27). For its use in the tissue-engineered cartilage, a 3D composite, such as a chondroitin sulfate A/chitosan hydrogel scaffold, was prepared. This hydrogel supported a differentiated phenotype of seeded articular chondrocytes and type II collagen and proteoglycan production (28). Also, the organic–inorganic hybrid scaffold, used as a chitosan/tricalcium phosphate scaffold, was fabricated for tissue-engineered bone. When osteoblast cells collected from rat fetal calvary were seeded onto a chitosan/tricalcium phosphate scaffold, the cells proliferated in a multiplayer manner and deposited a mineralized matrix (29).

Agarose. Agarose is another type of marine source polysaccharide purified extract from sea creatures, such as agar or agar-bearing algae. One of the unique properties of agarose is the formation of a thermally reversible gel, which starts to set at a concentration in excess of 0.1% at a temperature $\sim 40\,^{\circ}C$ and a gel melting temperature of $90\,^{\circ}C$. Agarose gel is widely used in the electrophoresis of proteins and nucleic acid. Its good gelling behavior could make it a suitable injectable bone substitute and cell

carrier matrix (17). Allogenic chondrocyte-seeded agarose gels have been used as a model to repair osteochondral defects in vivo. The repaired tissues were scored histologically based on the intensity and extent of the proteoglycan and the type II collagen immunoassay, the structural features of the various cartilaginous zones, integration with host cartilage, and the morphological features and arrangement of chondrocytic cells. The allogenic chondrocyte–agarose-grafted repairs had a higher semiquantitative score than control grafts. These results showed a good potential for use in tissue engineering (30). More detailed studies, such as the in vivo mechanical properties, biocompatibility and toxicity, and the balance degradation and synthesis kinetics of agarose-based tissue-engineered products, must be undertaken to further successful agarose applications (31).

Small Intestine Submucosa. Porcine small intestine submucosa (SIS) is an important material for natural ECM scaffolds (15). Many experiments have shown systematically that an acellular resorbable scaffold material, derived from SIS, is rapidly resorbed, supports early and abundant new blood vessel growth, and serves as a template for the constructive remodeling of several body tissues including musculoskeletal structures, skin, body wall, dura mater, urinary bladder, and blood vessels (32). The SIS material consists of a naturally occurring ECM, rich in components that support angiogenesis, including fibronectin, glycosaminoglycans including heparin, several collagens (including types I, III, IV, V, and VI), and angiogenic growth factors such as basic fibroblast growth factor and vascular endothelial cell growth factor (33). For these reasons, SIS scaffolds have been successfully used to reconstruct the urinary bladder, for vascular grafts, to reconstruct cartilage and bone alone or as a composite with synthetic polymers and inorganic biomaterials (34).

Acellular Dermis. Acellular human skin, that is skin removed of all cellular components, may be one of the most significant ECMs. An acellular dermis can be seeded with fibroblasts and keratinocytes to fabricate a dermal–epidermal composite for the regeneration of skin. AlloDerm (LifeCell, Branchburgh, NJ) is a typical commercialized product, a split-thickness acellular allograft prepared from human cadaver skin and cryopreserved for off-shelf use (35). Alloderm has been successful in the treatment of burn patients because of its nonantigenic dermal scaffold that includes elastin, proteoglycan, and basement membrane.

Poly(hydroxyalkanoates). Poly(hydroxyalkanoates) are entirely natural and are obtained from the microorganism *Alcaligen eutrophus* as Gram-negative bacteria. The physical properties of polyhydroxybutyrate (PHB) are similar to nondegradable polypropylene. Its copolymers with hydroxyvalerate [poly(hydroxybutylate-*co*-hydrovalerate); PHBV] have a modest range of mechanical properties and a correspondingly modest range of chemical compositions for monomers and processing conditions. Due to their good processability, these polymers can be manufactured into many forms, such as fibers, meshes, sponges, films, tubes, and matrices through standard processing techniques.

The family of poly(hydroxyalkanoates) does not appear to cause any acute inflammation, abscess formation, or tissue necrosis in whethers in the form of nonporous disks or cylinders, adjacent tissues (36). To optimize the mechanical property of PHBV, organic–inorganic hybrid composites such as PHBV–hydroxyapatite were developed for the tissue engineering of bone; hydroxyapatite promotes osteoconductive activity (13). Also, Schwann cell-seeded PHB was applied to regenerate a nerve in the shape of a conduit to guide and induce neonerve tissue at the nerve ends. Good nerve regeneration in PHB conduits as compared to nerve grafts was observed. The shape, mechanical strength, porosity, thickness, and degradation rate of PHB and its copolymers can be engineered.

Other Natural Polymers. Excluding those polymers discussed in the Natural Polymers section above, other natural polymers, are proteins, albumin, gluten, elastin, fibroin, cellulose, starch, sclerolucan, elsinan, pectin (pectinic acid), galactan, curdlan, gellan, levan, emulsan, dextran, pullulan, heparin, silk, and chondroitin 6-sulfate. Although they are not discussed here, these biopolymers are of interest because of their unusual and useful functional properties as well as their abundance. This group of natural polymers are (*1*) biocompatible and nontoxic, (*2*) easily processed as film and gel, (*3*) heat stable and thermal processable over a broad temperature range, and (*4*) water soluble (17). *In vivo* and *in vitro* experiments, and physicochemical modifications should be performed in the near future to promote the use of these natural polymers in tissue-engineered scaffolds.

Synthetic Polymers

Natural polymers are not used more extensively because they are expensive, differ from batch to batch, and there is a possibility of cross-contamination from unknown viruses or unwanted diseases due to their isolation from plant, animal, and human tissue. Alternatively, synthetic polymeric biomaterials have easily controlled physicochemical properties and quality, and no immunogenecity. Also, they can be processed by various techniques and supplied consistently in large quantities. To adjust the physical and mechanical properties of a tissue-engineered scaffold at a desired place in the human body, the molecular structure, and molecular weight are adjusted during the synthetic process. Synthetic polymers are largely divided two categories: biodegradable, and nonbiodegradable. Some nondegradable polymers include poly(vinylalcohol) (PVA), poly(hydroxylethylmethacryalte), and poly(*N*-isopropylacryamide). Some synthetic degradable polymers are in the family of poly(α-hydroxy ester)s, such as polyglycolide (PGA), polylactide (PLA) and its copolymer poly(lactide-*co*-glycolide) (PLGA), polyphosphazene, polyanhydride, poly(-propylene fumarate), polycyanoacrylate, polycaprolactone, polydioxanone and biodegradable polyurethanes.

Between these two polymers, synthetic biodegradable polymers are preferred for use in tissue-engineered scaffolds because they have minimal chronic foreign body reactions and they promote the formation of completely natural tissue. That is, they can form a temporary scaffold

for mechanical and biochemical support. More detailed polymer fabrication methods are discussed in the section, Scaffold Fabrication and Characterization.

Poly(α-Hydroxy Ester)s. The family of poly(α-hydroxy acid)s, such as PGA, PLA, and its copolymer PLGA, are among the few synthetic polymers approved for human clinical use by the U.S. Food and Drug Administration (FDA). These polymers are extensively used or tested as scaffold materials, because they are as bioerodible with good biocompatibility, have controllable biodegradability, and relatively good processability (37). This family of poly(α-hydroxy éster)s has been used for three decades: PGA as a suture; PLA in bone plate, screw and reinforced materials; and PLGA in surgical and drug delivery devices. The safety of these materials has been prove for many medical applications (38–47).

These polymers degrade by nonspecific hydrolytic scission of their ester bonds. Polyglycolide biodegrades by a combination of hydrolytic scission and enzymatic (esterase) action producing glycolic acid, which can either enter the tricarboxylic acid (TCA) cycle or be excreted in urine and eliminated as carbon dioxide and water. The hydrolysis of PLA yields lactic acid, which is a normal byproduct of anaerobic metabolism in the human body and is incorporated in the TCA cycle to be excreted finally by the body as carbon dioxide and water. With the addition of a methyl group to glycolide, PLA is much more hydrophobic than the highly crystalline PGA. As a result, PLA has a slower degradation rate over a year's time. The degradation time of PLGA as a copolymer can be controlled from weeks to over a year by varying the ratio of monomers, its molecular weight, and the processing conditions. The synthetic methods and physicochemical properties, such as melting temperature, glass transition temperature, tensile strength, Young's modulus, and elongation, were reviewed elsewhere (48).

The mechanism of biodegradation of poly(α-hydroxy acid)s is bulk degradation, which is characterized by a loss in polymer molecular weight, while its mass is maintained. Mass maintenance is useful for tissue-engineering applications that require a specific shape. However, a loss in molecular weight causes a significant decrease in mechanical properties. Degradation depends on its chemical history, porosity, crystallinity, steric hindrance, molecular weight, water uptake, and pH. Degradable products, such as lactic acid and glycolic acid, decrease the pH in the surrounding tissue resulting in inflammation and potentially poor tissue development. The PGA, PLA, and PLGA scaffolds are applied for the regeneration of all tissue, including skin, cartilage, blood vessel, nerve, liver, dura mater, bone, and other tissue (10,12,17). For the application of these polymers as scaffolds, the development of fabrication methods for porous structures is also important.

The hybrid structure of chondrocytes and fibroblast/PGA fiber felts was successfully tested in the regeneration of cartilage and skin, respectively (49). Also, porous PLGA scaffolds with an average pore sizes of 150–300 or 500–710 μm were seeded with osteoblast cells, which resulted in good bone generation. Composites of PLA/tricalcium phosphate and PLA/hydroxyapatite were attempted to induce bone formation both *in vitro* and *in vivo* (13,50). Porous PLA tubes with an inside diameter of 1.6 mm, are outside diameter of 3.2 mm, and lengths of 12 mm, were implanted into 12 mm gaps in the rat sciatic nerve model. Compared to control grafts, both the number and density of axons were significantly less for the tabulated implants. The PGA tube was also tested for the regeneration of vascular grafts, and showed good *in vivo* results.

To improve the physicochemical properties of poly(α-hydroxy acid)s for use as scaffold materials, the chemical modification of both end groups of PLA and PGA was undertaken; the additional reaction of the moieties helps to control the biological and/or physical properties of biomaterials (17). For example, poly(lactic acid-*co*-lysine-*co*-aspartic acid) (PLAL-ASP) was synthesized to add a cell adhesion property. Similarly, a copolymer of lactide and ε-caprolactone was synthesized to improve the elastic property of PLA. The PLA-poly(ethylene oxide) (PEO) copolymers were synthesized to have the degradative and mechanical properties of PLA and the biological control offered by PEO and its functionalization (51). One of the unique characteristics of PLA-PEO block copolymers is its temperature sensitivity. Because of the hydrophobicity of PLA and hydrophilicity of PEO, the sol–gel property can be applied to injectable cell carriers. Also, a nano-hybrid composite with other materials has been developed for application to all organs in the body.

Poly(vinyl Alcohol). Poly(vinyl alcohol) is synthesized from poly(vinyl acetate) by saponification. The result is a hydrogel that contains some water, which is similar to cartilage. It is relatively biocompatible, swells with a large amount of water, easily sterilized, and easily fabricated and molded into desired shapes. It has a reactive pendant alcohol group that can be modified by chemical cross-linking, physical cross-linking, or by incorporating an acrylate group, which results in improvement of its mechanical properties. A typical commercialized PVA gel is Salubria (Salumedica, Atlanta, GA), which was created by completing a series of freeze–thaw cycles with PVA polymers and 0.9% saline solution. By changing the ratio of PVA and H_2O, the molecular weight of PVA, and the quantity and duration of the freeze–thaw cycles, the physical properties of the PVA hydrogel can be controlled. Poly(vinyl alcohol) has been used in cartilage regeneration; it has similar mechanical properties needed in breast augmentation, diaphragm replacement, and bone replacement (10). One significant drawback is that it is not fully biodegradable because of the lack of labile bonds within the polymer backbone. So, it is recommended that low molecular weight PVA, $\sim 15{,}000\ \mathrm{g \cdot mol^{-1}}$, which can be absorbed through the kidney, might be applied to tissue-engineered scaffolds.

Polyanhydride. Polyanhydride is synthesized by the reaction of diacids with anhydride to form acetyl anhydride prepolymers. High molecular weight anhydrides are synthesized from the anhydride prepolymer in a melt condensation. Polyanhydrides are modified to increase their physical properties by a reaction with imides (17). A typical example of this is copolymerization with an

aromatic imide monomer that results in the polyanhydride-*co*-imide used in hard tissue engineering. To control degradability and to enhance mechanical properties, photo-crosslinkable functional groups were introduced by the substituted methacrylate groups on polyanhydrides for orthopedic tissue engineering (48,50). The degradation mechanism of polyanhydrides is a highly predictable and controlled, surface erosion whereas that of poly(α-hydroxy ester) is bulk erosion. To optimize the degradation behavior of anhydride-based copolymers, the polymer backbone chemistry needs to be controlled to achieve a ratio of monomer and molecular weight.

Poly(Propylene Fumarate). Poly(propylene fumarate) and its copolymer, a biodegradable and unsaturated linear polyester, were synthesized as potential scaffold biomaterials. The degradation mechanism is a hydrolytic chain scission similar to poly(α-hydroxy ester). The mechanical strength and degradable behaviors were controlled by crosslinking with a vinyl monomer at the unsaturated double bonds. The physical properties are enhanced by a composite with degradable bioceramic β-tricalcium phosphate, which is used as injectable bone (52). Copolymerization of propylene fumarate with ethylene glycol can be made elastic with poly(propylene fumarate) and used as a cardiovascular stent. New materials for propylene fumarate polymers are continually being investigated through copolymer synthesis, hybrid composites, and blends.

PEO and Its Derivatives. Poly(ethylene oxide) is one of the most important and widely used polymers in biomedical applications because of its excellent biocompatibility (51,53,54). It can be produced by anionic or cationic polymerization from ethylene oxide by initiators. Poly(ethylene oxide) is used to coat materials used in medical devices to prevent tissue and cell adhesion, as well as in the preparation of biologically relevant conjugates, and in induction cell membrane fusion. These PEO hydrogels can be fabricated by crosslinking reactions which gamma rays, electron beam irradiation, or chemical reactions. This hydrogel can be used for drug delivery and tissue engineering. Vigilon (C.R. Bard, Inc., Murray Hill, NJ) is a radiated crosslinked, high molecular weight PEO, which swells with water and is used as a wound-covering material. The hydroxyl in the glycol end group is very active, making it appropriate for chemical modification. The attachment of bioactive molecules, such as cytokines and peptides to PEO or poly(ethylene glycol) (PEG) promotes the efficient delivery of bioactive molecules. See the section, Cytokine Release System for Tissue Engineering, for a more detailed explanation.

To synthesize biodegradable PEO, block copolymerization with PGA or PLA degradable units has been carried out. The hydrogel can be polymerized into two- or three-block copolymers such as PEO-PLA, PEO-PLA-PEO, and PLA-PEO-PLA. For the biodegradable block, ε-caprolactone, δ-valerolactone, and PLGA can be used (50). A characteristic of this series of hydrogels is a temperature-sensitive phenomena. A solid state at room temperature changes to a gel state at body temperature. Hence, biodegradable hydrogels are very useful in injectable cell loading scaffolds (55). After injection of the chondrocyte cell hybrid structure and biodegradable hydrogels, the hydrogels degrade in vivo and neocartilage tissue remains.

Also, the copolymers of PEO and poly(propylene oxide) (PPO), including PPO-PEO-PPO or PEO-PPO-PEO block copolymers, are the basis for the commercially available Pluronics and Tetronics. Pluronics form a thermosensitive gel by shrinking hydrophobic segments of the copolymer PPO (54). The physicochemical property of the hydrogel can be varied with the composition and structure of the ratio of PPO and PEO. Some have been approved by the FDA and EPA for use as food additives, pharmaceutical ingredients, and agricultural products. Although the polymer is not degraded by the body, the gels dissolve slowly and the polymer is eventually cleared. Chondrocytes-loaded Pluronics, when directly injected at the injured site containing tissue-engineered cartilage, maintained its original shape in the developing neocartilage (56). Also, these polymers are used in the treatment of burn patients and for protein delivery. The advantages of these injectable hydrogels include: (*1*) no need for surgical intervention, (*2*) easy pore-size manipulation, and (*3*) no need for complex shape fobrication.

Polyphosphazene. Polyphossphazene consists of an inorganic backbone of alternating single and double bonds between phosphorous and nitrogen atoms, while most of the polymer is made up of a carbon–carbon organic backbone (10,12,17). It has side groups that can react with other functional groups which result in block or star polymers. Biological and physical properties can be controlled by the substitution of functional side groups. For example, the rate of degradation can be varied by controlling the proportion of hydrolytically labile side groups. The wettability such as hydrophilicity, hydrophobicity, and amphiphilicity, of polyphosphazene might be dependent on the properties of the side group. It can be made into films, membranes, and hydrogels for scaffold applications by cross-linking or grafting modifications (48). The cytocompatibility of highly porous polyphosphazene scaffolds offers possibilities for skeletal tissue engineering. Also, the blend of polyphosphazene with PLGA may be modified and its miscibility and degradability determined (57).

Biodegradable Polyurethane. Polyurethane is one of the most widely used polymeric biomaterials in biomedical fields due to its unique physical properties, such as durability, elasticity, elastomer-like character, fatigue resistance, compliance, and tolerance. Moreover, the reactivity of the functional group of the polyurethane backbone can be achieved by the attachment of biologically active biomolecules and the adjustment of their hydrophilicity–hydrophobicity (58). Recently, the synthesis of a new generation of nontoxic biodegradable peptide-based polyurethanes was achieved. Typical biodegradable polyurethane is composed of an amino acid-based hard segment (such as lysine diisocyanate), a polyol soft segment (such as a hydroxyl doner-like polyester), and sugar (59). Hence, the degradation products of these nontoxic lysine diisocyanate-based urethane polymers are nontoxic lysine and the polyol. If the covalent bonding of various proteins, such

as cytokines, growth factors, and peptides, are introduced in the polymer backbone, the controlled release of the bioactive molecules can be achieved in a degradable manner using polyurethane scaffolds. The mechanisms of degradation are hydrolysis, oxidation—both thermal, and enzymatic. Both the chemistry and the composition of soft and hard segments play an important role in the degradability of polyurethane. Poly(urethane-urea) matrices with lysine diisocyanate as the hard segment and glucose, glycerol, or PEG as the soft segments have been studied. In the application of biodegradable polyurethane as a scaffold various types of cells, such as chondrocytes, bone marrow stromal cells, endothelial cells, and osteoblast cells, were successfully adhered and proliferated. Also, toxicity, induction of a foreign body reaction, and antibody formation were not observed in in vivo experiments. The long-term safety and biocompatibility of biodegradable polyurethane must be continuously monitored for use in tissue-engineered scaffold substrates.

Other Synthetic Polymers. Besides the synthetic polymers already introduced in the above sections, many other synthetic polymers, either degradable or nondegradable, are being developed and tested to mimic the natural tissue and wound-healing environment. Examples are poly(2-hydroxyethylmethacrylate) hydrogel, injectable poly(*N*-isopropylacrylamide) hydrogel, and polyethylene for neocartilage; poly(iminocarbonates) and tyrosine-based poly(iminocarbonates) for bone and cornea; crosslinked collagen–PVA films and an injectable biphasic calcium phosphate–methylhydroxypropylcellulose composite for bone regeneration materials; a polyethylene oxide-*co*-polybutylene terephthalate for bone bonding; poly(orthoester) and its composites with ceramics for tissue-engineered bone; synthesized conducting polymer polypyrrole–hyaruronic acid composite films for the stimulation of nerve regeneration; and peptide-modified synthetic polymers for the stimulation of cell and tissue.

It is very important for the design and synthesis of more biodegradable and biocompatible scaffold biomaterials to mimic the natural ECM in terms of bioactivity, mechanical properties, and structures. The more biocompatible biomaterials tend to elicit less of an immune response and reduce an inflammatory response at the implantation site.

Bioceramic Scaffolds

Bioceramic is a term used for biomaterials that are produced by sintering or melting inorganic raw materials to create an amorphous or a crystalline solid body that can be used as an implant. Porous final products have been used mainly as scaffolds. The components of ceramics are calcium, silica, phosphorous, magnesium, potassium, and sodium. Bioceramics used in tissue engineering might be classified as nonresorbable (relatively inert), bioactive, or surface active (semi-inert), and biodegradable or resorbable (non-inert). Alumina, zirconia, silicone nitride, and carbons are inert bioceramics. Certain glass ceramics are dense hydroxyapatites [$9CaO \cdot Ca(OH)_2 \cdot 3P_2O_5$] and semi-inert (bioactive). Calcium phosphates, aluminum–calcium–phosphates, coralline, tricalcium phosphates ($3CaO \cdot P_2O_5$), zinc-calcium-phosphorous oxides, zinc-sulfate-calcium-phosphates, ferric–calcium–phosphorous–oxides, and calcium aluminates are resorbable ceramics (60). Among these bioceramics, synthetic apatite and calcium phosphate minerals, coral-derived apatite, bioactive glass, and demineralized bone particles are widely used in the hard tissue engineering area, hence, they will be discussed in this section.

Synthetic crystalline calcium phosphate can be crystallized into salts such as hydroxyapatite and β-whitlockite, depending on the Ca/P ratio. These salts are very tissue compatible and are used as bone substitutes in a granular, sponge form or as a solid block. The apatite formed with calcium phosphate is considered to be closely related to the mineral phase of bone and teeth. The chemical composition of crystalline calcium phosphate is a mixture of $3CaO \cdot P_2O_5$, $9CaO \cdot Ca(OH)_2 \cdot 3P_2O_5$ and calcium pyrophosphate ($4CaO \cdot P_2O_5$). The active exchange of ions occurs on the surface and leads to the exchange composition of minerals (9,61). When porous ceramic scaffolds were implanted in the body, both with or without cells for tissue-engineered bone, the delivery of some elements to the new bone was at the interface between the materials and the osteogenic cells.

Tricalcium phosphate is the rapidly resorbable calcium phosphate ceramic resulting in resorption 10–20 times faster than hydroxyapatite (13). Porous tricalcium phosphate may stimulate local osteoblasts for new bone formation. Injectable calcium phosphate cement containing β-tricalcium phosphate, dibasic dicalcium phosphate, and tricalcium phosphate monoxide, was investigated for the treatment of distal radius fractures. Calcium sulfate hemihydrate (plaster of Paris), as a synthetic graft material, was also tested for tissue-engineered bone.

Coral-derived apatite (Interpore; Interpore international, Irvine, CA) is a natural substance made by marine vertebrate (62). The porous structure of coral, which is structurally similar to bone, is a unique physicochemical property that promotes its use as a scaffold matrix for bone. The main component of natural coral is calcium carbonate or aragonite, the metastable form of calcium carbonate. This compound can be converted to hydroxyapatite by a hydrothermal exchange process, which results in a mixture of hydroxyapatites, $9CaO \cdot Ca(OH)_2 \cdot 3P_2O_5$, and fluoroapatite, $Ca_5(PO_4)_3F$. For tissue-engineered bone, the hybrid structure of porous coral-derived scaffolds and mesenchymal stem cells were demonstrated in vitro. The results showed the differentiation of bone marrow derived from stem cells to osteoblasts; successive mineralizations were successfully accomplished (63).

Glass ceramics are polycrystalline materials manufactured by controlled crystallization of glasses using nucleating agents, such as small amounts of metallic agent Pt groups, TiO_2, ZrO_2, and P_2O_5, which result in a fine-grained ceramic that possesses excellent mechanical and thermal properties (60,61). Typical bioglass ceramics developed for implantations are SiO_2-CaO-Na_2O-P_2O_5 and Li_2O-ZnO-SiO_2 systems. These bioglass scaffolds are suitable for inducing direct bonding with bone. Bonding to bone is related to the composition of each component.

One significant natural bioactive material is the demineralized bone particle, which is a powerful inducer of new

bone growth (38,41). Demineralized bone particles contain many kinds of osteogenic and chondrogenic cytokines such as bone morphogenetic protein, and are widely used as filling agent for bony defects. Because of their improved availability through the tissue bank industry, demineralized bone particles are widely used in clinical settings. To achieve more optimal results in the application of demineralized bone particles to tissue engineering, nanohybridization with synthetic (PLGA/demineralized bone particle hybrid scaffolds) and with natural organic compounds (collagen/demineralized bone particle hybrid scaffolds), has been carried out.

Porosity—the size of the mean diameter and the surface area—is a critical factor for the growth and migration of tissue into bioceramic scaffolds (60). Several methods have been introduced to optimize the fabrication of porous ceramics, such as dip casting, starch consolidation, the polymeric sponge method, the foaming method, organic additives, gel casting, slip casting, direct coagulation consolidation, hydrolysis-assisted solidification, and freezing methods. Therefore, it is very important to choose an appropriate method of preparation based on the physical properties of the desired organs.

A CYTOKINE-RELEASE SYSTEM FOR TISSUE ENGINEERING

Growth factors, a type of cytokine, are polypeptides that transmit signals to modulate cellular activity and tissue development including cell patterning, motility, proliferation, aggregation, and gene expression. As in the development of tissue-engineered organs, regeneration of functional tissue requires maintenance of cell viability and differentiated function, encouragement of cell proliferation, modulation of the direction and speed of cell migration, and regulation of cellular adhesion. For example, transforming growth factor-β_1 (TGF-β_1) might be required to induce osteogenesis and chondrogenesis from bone marrow derived mesenchymal stem cells. Also, brain-derived neurotrophic factor (BDNF) can be enhanced to regenerate the spinal cord after injury. The easiest method for the delivery of growth factor is injection near the site of cell differentiation and proliferation (4). The most significant problems associated with the direct injection method are that the growth factors have a relatively short half-life, have a relatively high molecular weight and size, display very low tissue penetration, and have potential toxicity at systemic levels (4,10,11,16).

A promising technique for the improvement of their efficacy is to locally control the release of bioactive molecules for a specified release period to promote impregnation into a biomaterial scaffold. Through impregnation into the scaffold carrier, protein structure and biological activity can be stabilized to a certain extent, resulting in prolonging the release time at the local site. The duration of cytokine release from a scaffold can be controlled by the types of biomaterials used, the loading amount of cytokine, the formulation factors, and the fabrication process. The release mechanisms are largely divided into three categories: (*1*) diffusion controlled, (*2*) degradation controlled, and (*3*) solvent controlled. The mechanism of biodegradable scaffold materials was regulated by degradation control, whereas that of the nondegradable material was regulated by diffusion and/or solvent control. The desired release pattern, such as a constant, pulsatile, and time programed behavior over the specific site and injury can be achieved by the appropriate combination of these mechanisms. Also, the cytokine-release system's geometries and configurations can be altered to produce the necessary scaffold, tube, microsphere, injectable form or fiber (46,51,54).

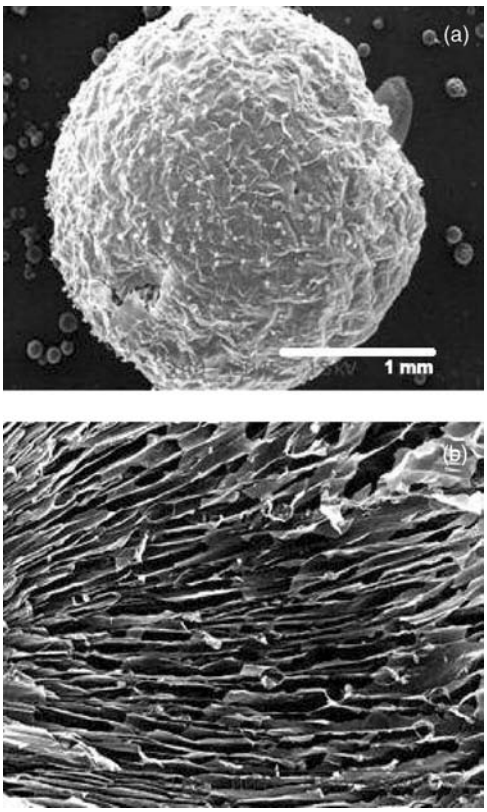


Figure 4. (a) Bone marrow-derived mesenchymal stem cells impregnated TGF-β_1 loaded alginate beads (original magnifications 40×), and (b) inner structure of alginate beads (original magnifications 100×).

Figures 4–6 show the TGF-β_1 loaded alginate bead and the release pattern of TGF-β_1 from alginate beads for the chondrogenesis from bone marrow-derived mesenchymal stem cells (64). The pore structure of 10 μm width and 100 μm length, was well suited to promote cell proliferation (Fig. 4); TGF-β_1 released at a near zero-order rate for 35 days (Fig. 5). By using the alginate bead with TGF-β_1 delivery system, chondrogenesis was successfully attained, as shown in Fig. 6.

To fabricate a new sustained delivery device for nerve growth factor (NGF), we developed NGF-loaded biodegradable PLGA films by a novel and simple sandwich solvent casting method for possible applications in the central nervous system (45). The release of NGF from the NGF-loaded PLGA films was prolonged $>$ 35 days with a zero-order rate, without initial burst, and controlled by variation of different molecular weights and different NGF loading amounts as shown in Fig. 7. After 7 days, NGF

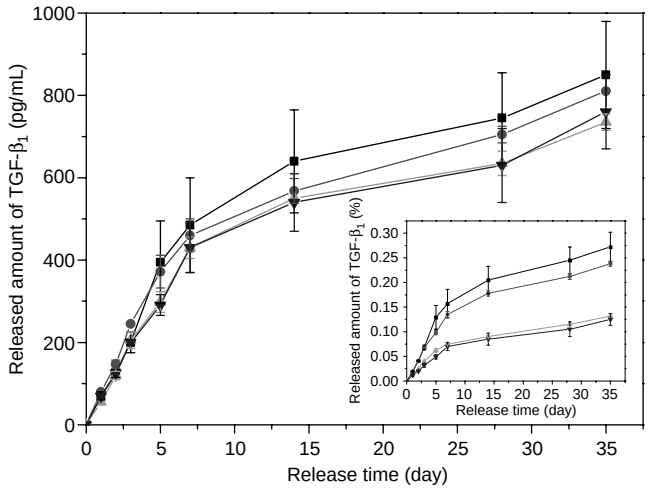


Figure 5. Release pattern of TGF-β_1 from TGF-β_1 loaded alginate beads; (■) 0.5 μg TGF-β_1, (●) 0.5 μg TGF-β_1 with heparin, (▲) 1.0 μg TGF-β_1, and (▼) 1.0 μg TGF-β_1 with heparin.

was released in a phosphate buffered saline solution (PBS; pH 7.0) and rat pheochromocytoma (PC-12) cells were cultured on the NGF-loaded PLGA film for 3 days. The released NGF stimulated neurite sprouting in the cultured

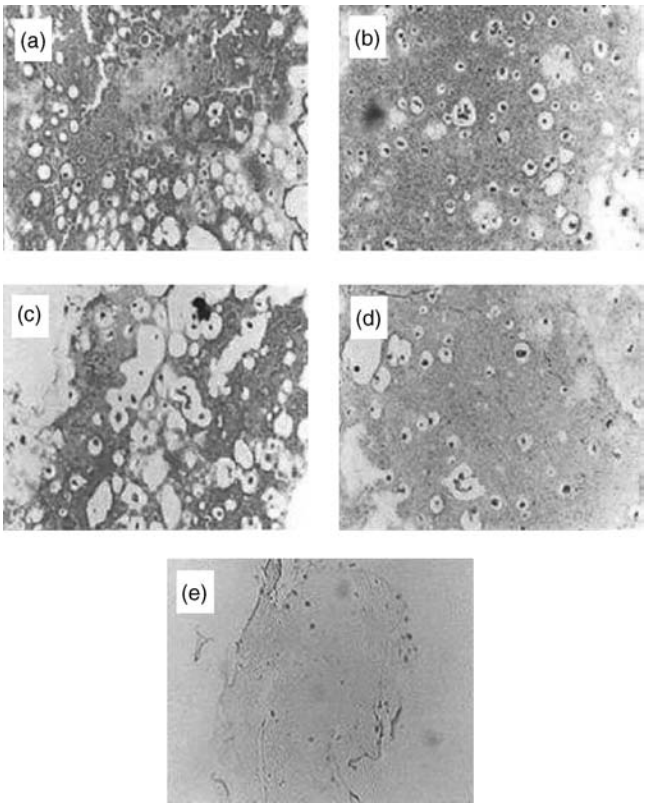


Figure 6. Safranin-O staining of chondrogenesis cells from bone marrow-derived mesenchymal stemcells in alginate beads. We can observe typical chondrocyte cells in alginate beads; (a) 0.5 μg·mL^{-1} TGF-β_1, (b) 1.0 μg·mL^{-1} TGF-β_1, (c) 0.5 μg·mL^{-1} TGF-β_1 with heparin (d) 1.0 μg·mL^{-1} TGF-β_1 with heparin, and (e) control (without TGF-β_1) (Original magnification 100×).

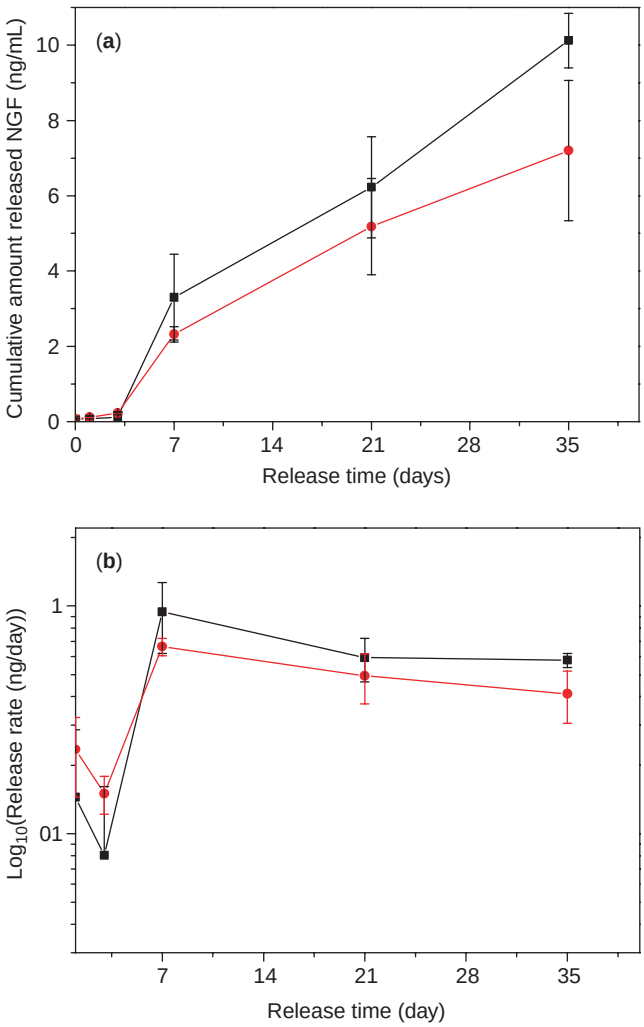


Figure 7. (a) Release profiles and (b) logarithmic plot of release rate for NGF from NGF-loaded PLGA films of 43,000 g/mol. (●) 25.4 ng, and (■) 50.9 ng NGF/cm^2 PLGA.

PC-12 cells; the remaining NGF in the NGF/PLGA film at 37°C for 7 days was still bioactive, as shown in Fig. 8. These studies suggest that NGF-loaded PLGA sandwich film can be released in the delivery system over the desired time period, thus, it can be a useful neuronal growth culture serving as a nerve contact guidance tube for applications in neural tissue engineering.

One serious problem during the fabrication of cytokine-loaded scaffolds is the denaturation and deactivation of cytokines, which result in loss of biological activity (65,66). Hence, the optimized method must be developed for stabilized cytokine-release scaffolds. For example, the release of NGF from a PLGA matrix was investigated using codispersants, such as polysaccharides (dextran) and proteins (albumin and β-lactoglobulin), with different molecular weights and charges. Negatively charged codispersants stabilized NGF in the PLGA system. Similarly, albumin stabilized epidermal growth factor (EGF) and heparin stabilized other growth factors.

Another available emerging technology is the "tethering of protein", that is, immobilization of protein on the surface

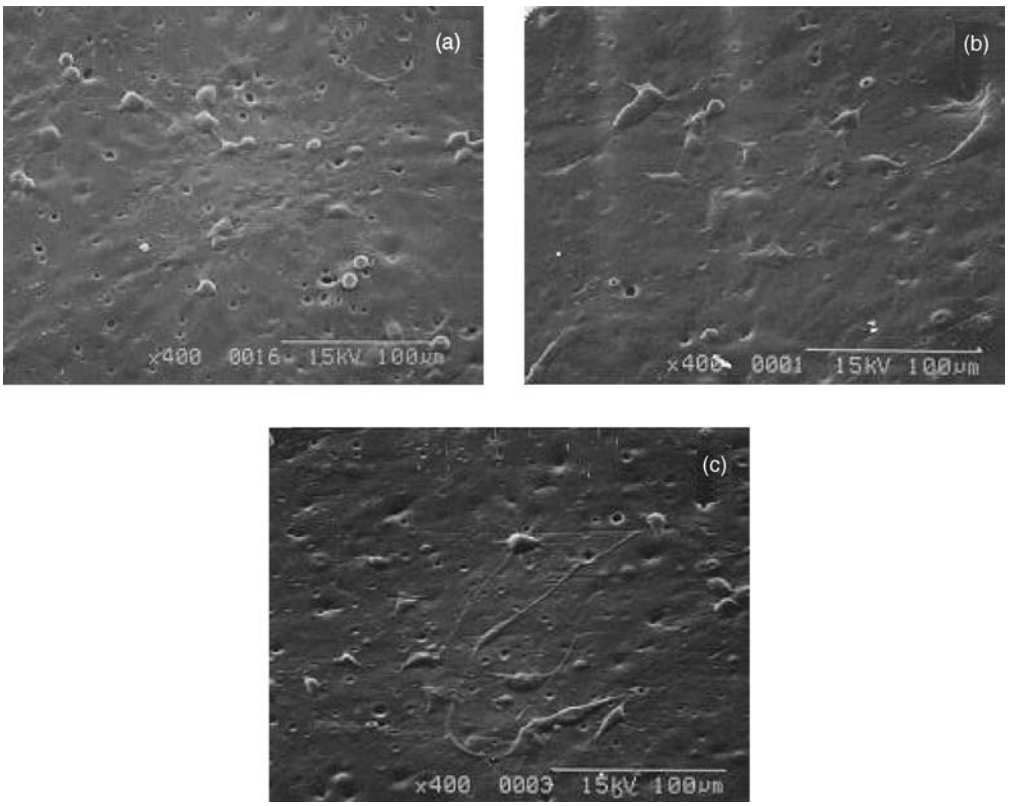


Figure 8. Effect of NGF released on neurites formation of PC-12 cells for 3-day cultivation on control (a) PLGA, (b) 25.4 ng, and (c) 50.9 ng NGF/cm^2 PLGA just after 7 days. There were total medium changes (Molecular weight of PLGA; 83,000 g/mol, original magnification; 400×).

of a scaffold matrix. Immobilization of insulin and transferrin to the poly(methylmetacrylate) films stimulates the growth of fibroblast cells compared to the same concentrations of soluble or physically adsorbed proteins (67). For the enhancement of cytokine activity, the PEO chain was applied as a short spacer between the surface of the scaffold and the cytokine. Tethered EGF, immobilized to the scaffold through the PEO chain, showed better DNA synthesis or cell rounding compared to the physically adsorbed EGF surface (68).

Conjugation of cytokine with an inert carrier prolongs the short half-life of protein molecules. Inert carriers are albumin, gelatin, dextran, and PEG. In PEGylation, PEG conjugated cytokine is most widely used for the release. This carrier appears to decrease the rate of cytokine degradation, attenuate the immunological response, and reduce clearance by the kidneys (69). Also, this PEGylated cytokine can be impregnated into scaffold materials by physical entrapment for sustained release. For example, the NGF-conjugated dextran (70,000 $g \cdot mol^{-1}$) impregnated polymeric device was implanted directly into the brain of adult rats. Conjugated NGF could penetrate into the brain tissue 8 times faster than the unconjugated NGF. This conjugation method can be applied to the delivery of proteins and peptides. Immobilized RGD (arginin-glycine-aspartic acid) and YIGSR (tyrosin-leucineglycine-serine-arginine), which are typical ECM proteins, can enhance cell viability, function, and recombinant products in the cell (70).

Gene-activating scaffolds are being designed to deliver the targeted gene that results in the stimulation of specific cellular responses at the molecular level (4,3,11). Modification of bioactive molecules with resorbable biomaterial systems obtain specific interactions with cell integrins resulting in cell activation. These bioactive bioglasses and macroporous scaffolds also can be designed to activate genes that stimulate regeneration of living tissue (9). Gene delivery would be accomplished by complexation with positively charged polymers, encapsulation, and gel by means of the scaffold structure (51). Methods of gene delivery for gene-activating scaffolds are almost the same methods as for those with protein, drug, and peptides.

SCAFFOLD FABRICATION AND CHARACTERIZATION

Scaffold Fabrication Methods

Engineered scaffolds may enhance the functionalities of cells and tissues to support the adhesion and growth of a large number of cells because they provide a large surface area and pore structure within a 3D structure. The pore structure needs to provide enough space, permit cell suspension, and allow penetration of the 3D structure. Also, these porous structures help to promote ECM production, transport nutrients from nutrient media, and excrete waste products (10,12,15). Therefore, an adequate pore size and a uniformly distributed, and an interconnected pore structure, which allow for easy distribution of cells

throughout the scaffold structure, are very important. Scaffold structures are directly related to their fabrication methods; over 20 methods have been proposed (10,71).

The most common and commercialized scaffold is the PGA nonwoven sheet (Albany International Research Co., Mansfield, MA; porosity $\sim 97\%$, ~ 1–5 mm thick); it is one of the most tested scaffolds for tissue-engineered organs. To stabilize dimensionally and provide mechanical integrity, fiber-bonding technology was developed using heat and PLGA or PLA solution spray coating methods (72).

Porogen leaching methods have been combined with polymerization, solvent casting, gas foaming, or compression molding of natural and synthetic scaffolds biomaterials. The leaching of pore-generating particles such as sodium chloride crystal, sodium tartrate, and sodium citrate were sieved using a molecular sieve (10,71). PLGA, PLA, collagen, poly(orthoester), or SIS-impregnated PLGA scaffolds were successfully fabricated into a biodegradable sponge structure by this method with $> 93\%$ porosity and a desired pore size of 1000 μm. By using the solvent casting/particulate leaching method, complex geometries, such as tube, nose, and specific organ types (e.g., nano-composite hybrid scaffolds), could be fabricated by means of conventional polymer-processing techniques, such as calendaring, extrusion, and injection. Complex geometry can be fabricated from porous film lamination (33,39,42,47). The advantage of this method is its easy control of porosity and geometry. However, the disadvantages include: (*1*) the loss of water-soluble biomolecules or cytokines during the leaching porogen process, (*2*) the possibility that the remaining porogen as a salt can be harmful to the cell culture, and (*3*) the different geometry surface and cross-section that results.

The gas-foaming method consists of a solid scaffold matrix exposured to a sudden expansion of CO_2 gas under high pressure, which results in the formation of a sponge structure due to nucleation and expansion in a dissolved CO_2 scaffold matrix. The PLGA scaffolds with $> 93\%$ porosity and ~ 100 μm median pore size were developed by this method (71). A significant advantage is that there is no loss of bioactive molecules in the scaffold matrix, since there is no more need for the leaching process and there is no residual organic solvent. The disadvantage is the presence of a skimming film layer on the scaffold surface, which results in a need for an additional process to remove the skin layer.

The phase-separation method is divided into the freeze–drying, freeze–thaw, freeze–immersion precipitation, and emulsion freeze–drying techniques (37,72,73). Phase separation by freeze–drying can be induced by the appropriate concentration of polymer solution obtained by rapid freezing. Then, the used solvent is removed by freeze–drying, leaving in porous structure made up of a portion of the solvent. These can be collagen scaffolds with pores ~ 50–150 μm; collagen–glycosaminoglycan blend scaffolds with an average pore size ~ 90–120 μm; or chitosan scaffolds with a pore size ~ 1–250 μm, dependent on the freezing conditions (71). Also, scaffold structures of synthetic polymers, such as PLA or PLGA, have been successfully made much $> 90\%$ porosity and ~ 15–250 μm size by this method. The freeze–thaw technique induces phase separation between a solvent and a hydrophilic monomer upon freezing, followed by the polymerization of the hydrophilic monomer by means of ultraviolet (UV) irradiation and removal of the solvent by thawing. This technique leads to the formation of a macroporous hydrogel. A similar method is the freeze–immersion precipitation technique. The polymer solution is cooled, immersed in a nonsolvent, and then the vaporized solvent leads to a porous scaffold structure. Also, the emulsion freeze–drying method is used to fabricate a porous structure. Mixtures of polymer solution and nonsolvent are thoroughly sonicated, freezed quickly in liquid nitrogen at $-198\,^{\circ}C$, and then freeze–dried, resulting in a sponge structure. The advantage of these techniques is that they result in the loading of hydrophilic or hydrophobic bioactive molecules, whereas the disadvantages are relatively small pore sized scaffolds with precise pore structures that are hard to control (73).

Nano-electrospinning of PGA, PLA, PLGA, caprolactone copolymers, collagen, and elastin, has been extensively developed (74). For example, electrostatic processing can consistently produce PGA fiber diameters $\leqslant 1$ μm. By controlling the pick-up of these fibers, the orientation and mechanical properties can be tailored to the specific needs of the injured site. Also, collagen electrospinning was performed utilizing type I collagen dissolved in 1,1,1,3,3,3-hexafluoro-2-propanol with a concentration of $0.083\ g \cdot mL^{-1}$. The optimally electrospun type I collagen nonwoven fabric appeared with an average diameter of 100 ± 40 nm, which resulted in biomimicking fibrous scaffolds.

Injectable gel scaffolds have also been reported (10,16,51,54). An injectable, gelforming scaffold offers several advantages: (*1*) it can fill any space based on its ability to flow; (*2*) it can load various types of bioactive molecules and cells by simple mixing; (*3*) it does not contain residual solvents that may be present in a preformed scaffold; and (*4*) it does not require a surgical procedure for placement. Typical examples are thermosensitive gels such as Pluronics and PEG-PLGA-PEG triblock copolymer, pH sensitive gels such as chitosan and its derivates, an ionically cross-linked gel such as alginate, and fibrin and hyaluronan gels, as well as others previously introduced in the Natural Polymers section. In the near future, multifunctional gels which are tissue-specific, have a very fast sol–gel transition, are fully degradable over the necessary time period will be available.

Newly hybridized fabrication techniques such as organic–inorganic and synthetic–natural techniques at the nanosize level that biomimic, are also being developed for use in engineered scaffolds.

Physicochemical Characterization of Scaffolds

For the successful achievement of 3D scaffolds, several characterization methods are needed. These methods can be divided into four categories. (*1*) Morphology—porosity, pore size, and surface area; (*2*) mechanical properties—compressive and tensile strength; (*3*) bulk properties—degradation and its relevant mechanical properties; and (*4*) surface properties—surface energy, chemistry, and charge.

Porosity is defined as the fraction of the total volume occupied by voids that appear as percentages. The most widely used methods for the measurement of porosity are mercury porosimetry, scanning electron microscopy (SEM), and confocal laser microscopy.

Mechanical properties are extremely important when designing tissue-engineered products. Conventional testing instruments can be used to determine the mechanical properties of a porous structure. Mechanical tests can be divided into (*1*) creep tests, (*2*) stress–relaxation tests, (*3*) stress–strain tests, and (*4*) dynamic mechanical tests. These test methods are similar to those used for conventional biomaterials.

The rate of degradation of manufactured scaffolds is a very important factor in the design of tissue-engineered products. Ideally, the scaffold constructs provide mechanical and biochemical supports until the entire tissue regenerates, then the scaffold completely biodegrades at a rate consistent with tissue generation. Immersion studies are commonly conducted to track the degradation of the biodegradable matrix. Changes in weight loss and molecular weight can be evaluated by the chemical balance of the matrix, by SEM, and by gel permeation chromatography. These results produce the mechanism of biodegradation.

It is generally recognized that the adhesion and proliferation of different types of cells on polymeric materials depend largely on the materials' surface characteristics, such as wettability (hydrophilicity/hydrophobicity of surface free energy), chemistry, charge, roughness, and rigidity (37,40,41,44,45). The 3D aspects of tissue engineering are more important for cell migration, proliferation, DNA/RNA synthesis, and phenotype presentation on the scaffold materials. Surface chemistry and charge can be analyzed by electron scanning chemical analysis and streaming potential, respectively. Also, wettability of the scaffold surface can be measured by the contact angle using static and dynamic methods.

SURFACE MODIFICATION OF SCAFFOLDS FOR THE IMPROVEMENT OF BIOCOMPATIBILITY

As explained above, the surface properties of scaffold materials are very important. For example, the hydrophobic surfaces of PLA, PGA, and PLGA possess high interfacial free energy in aqueous solutions, which tends to unfavorably influence their cell, tissue, and blood compatibility in the initial stage of contact. Moreover, it does not allow the nutrient media to permeate into the center of the scaffolds. For these reasons, a surface treatment is applied by several methods: (*1*) chemical treatment using oxidants, (*2*) physical treatment using glow discharge, and (*3*) a blend with hydrophilic biomaterials or bioactive molecules.

The physicochemical treatment has been demonstrated to improve the wetting property and hydrophilicity of PLGA porous scaffolds fabricated by the emulsion freeze–drying method (37,45). The chemical treatments were 70% perchloric acid, 50% sulfuric acid, and 0.5 *N* sodium hydroxide solution. The physical methods included corona and plasma treatments generated by a radiofrequency glow discharge. After treatment, water contact angles decreased (Fig. 9). The wetting property of chemically treated PLGA scaffolds also ranked in the order of perchloric acid, sulfuric acid, and sodium hydroxide solution by blue dye intrusion experiment, whereas physical methods had no effect, as shown in Fig. 10. Thus, the chemical treatment method may be useful in uniform cell seeding into porous biodegradable PLGA scaffolds. Wettability plays an important role in cell adhesion, spreading, and growth on the PLGA surface, and the intrusion of nutrient media into the PLGA scaffold.

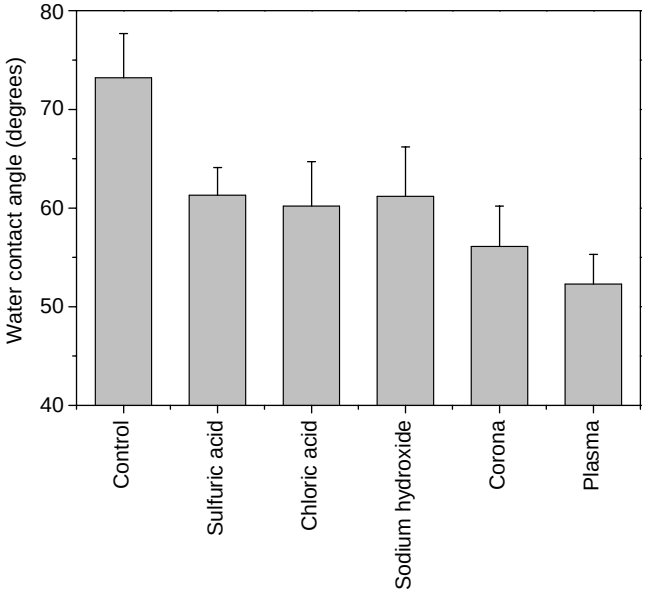


Figure 9. Changes of water contact angles after physicochemical treatment. The significant decreasing of water contact angle, that is, increased hydrophilicity, was observed.

Scaffolds impregnated with bioactive and hydrophilic material might be better for cell proliferation, differentiation, and migration due to cell stimulation. To give scaffolds new bioactive functionality from SIS powder as a natural source, scaffolds consisting of porous SIS/PLA and SIS/PLGA as a natural–synthetic composite, were prepared by the solvent casting–salt leaching method for use in tissue-engineered bone. A uniform distribution of good interconnected pores from the surface-to-core region was observed (pore size 40–500 μm), independent of the SIS amount, by using the solvent casting–salt leaching method. Porosities, specific pore areas as well as pore size distribution were also similar. After the fabrication of SIS/PLGA hybrid scaffolds, the wetting properties were greatly improved resulting in more uniform cell seeding and distribution, as shown in Fig. 11. Five different scaffolds, a PGA nonwoven mesh scaffold without glutaraldehyde (GA) treatment, PLA scaffolds without and with GA treatment, PLA/SIS scaffolds without and with GA treatment, were implanted into the back of nude mouse to observe the effect of SIS on the induction of cell proliferation by hematoxylin and eosin using von Kossa staining, for 8 weeks. It was observed that the effect of PLA/SIS scaffolds with GA treatment on bone induction is stronger than PLA scaffolds, that is the effects of PLA/SIS scaffolds with GA treatment $>$ PLA/SIS scaffolds without GA treatment $>$ PGA nonwoven $>$ PLA scaffolds only with GA treatment $=$ PLA scaffolds only without GA treatment for osteoinduction activity (Fig. 12).

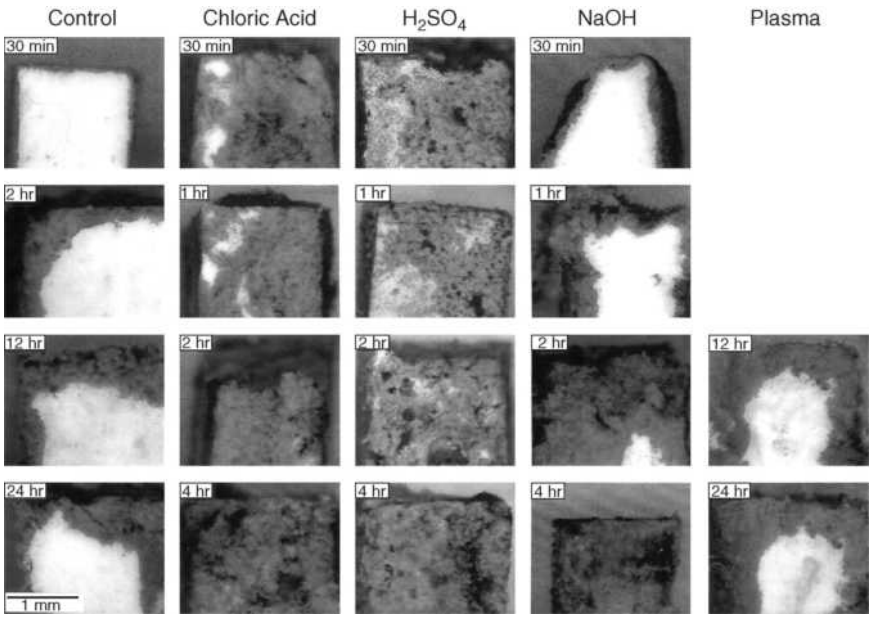


Figure 10. Wetting properties of physicochemically treated porous PLGA scaffolds by blue dye intrusion methods for 0.5, 1, 2, 4, 12, and 24 h.

STERILIZATION METHODS FOR SCAFFOLDS

The sterilizability of polymeric scaffold biomaterials is an important property, since polymers have lower thermal and chemical stability than other materials, such as ceramics and metals. Consequently, polymers are more difficult to sterilize using conventional techniques. Commonly used sterilization techniques are dry heat, autoclaving, radiation, and ethylene oxide gas (EOG). In addition, plasma glow discharge and electron beam sterilization recently were proposed due to their convenience (6,75).

In dry heat sterilization, the temperature varies between 160 and 190 °C. This temperature is above the melting and softening temperatures of many linear polymers, such as PLGA, resulting in the shrinking of the scaffold dimension. The PLA scaffolds were sterilized at 129 °C for 60 s, resulting in a minimal change in tensile properties. One of the significant problems was a decrease in molecular weight, which might have an affect on the degradation kinetics of the polymers. In the case of polyamide (Nylon) used as a nonbiodegradable polymer, oxidation occurs at the dry sterilization temperature, even though this is below its melting temperature. The only polymers that can safely be dry sterilized are polytetrafluoroethylene (PTFE) and silicone rubber. However, ceramic and metallic scaffolds were safe in this temperature range.

Steam sterilization (autoclaving) is performed under high steam pressure at a relatively low temperature (125–130 °C). However, if the polymer is subjected to attack by water vapor, this method cannot be employed. The PVC, polyacetals, PE (low density variety), and polyamides belong to this category. In the poly(α-hydroxy ester) family, a trace of water can deteriorate the PLGA backbone.

Chemical agents such as EOG and propylene oxide gases, and phenolic and hypochloride solutions are used widely for sterilizing all biomaterials, since they can be used at relatively low temperatures. Chemical agents

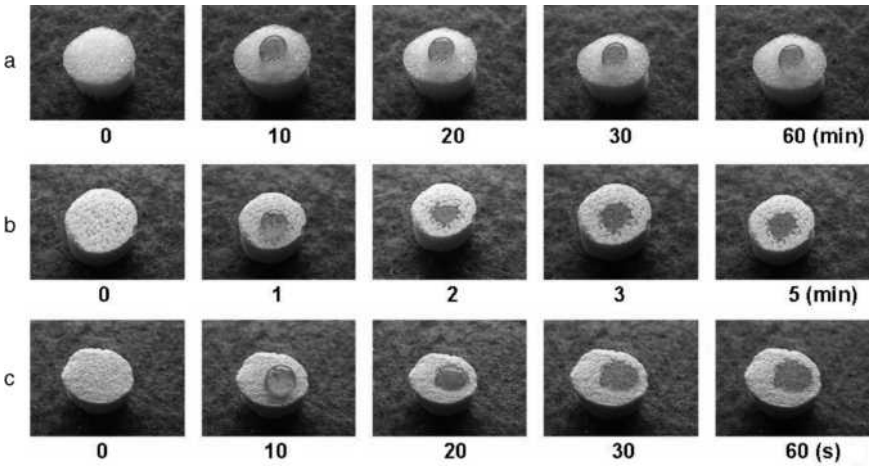


Figure 11. Wetting properties of SIS impregnated PLGA scaffolds by red dye intrusion methods. We observed the rapid penetration of water into SIS/PLGA scaffolds compared to the control PLGA scaffolds; (a) control PLGA, (b) 40% SIS/PLGA, and (c) 160% SIS/PLGA scaffolds.

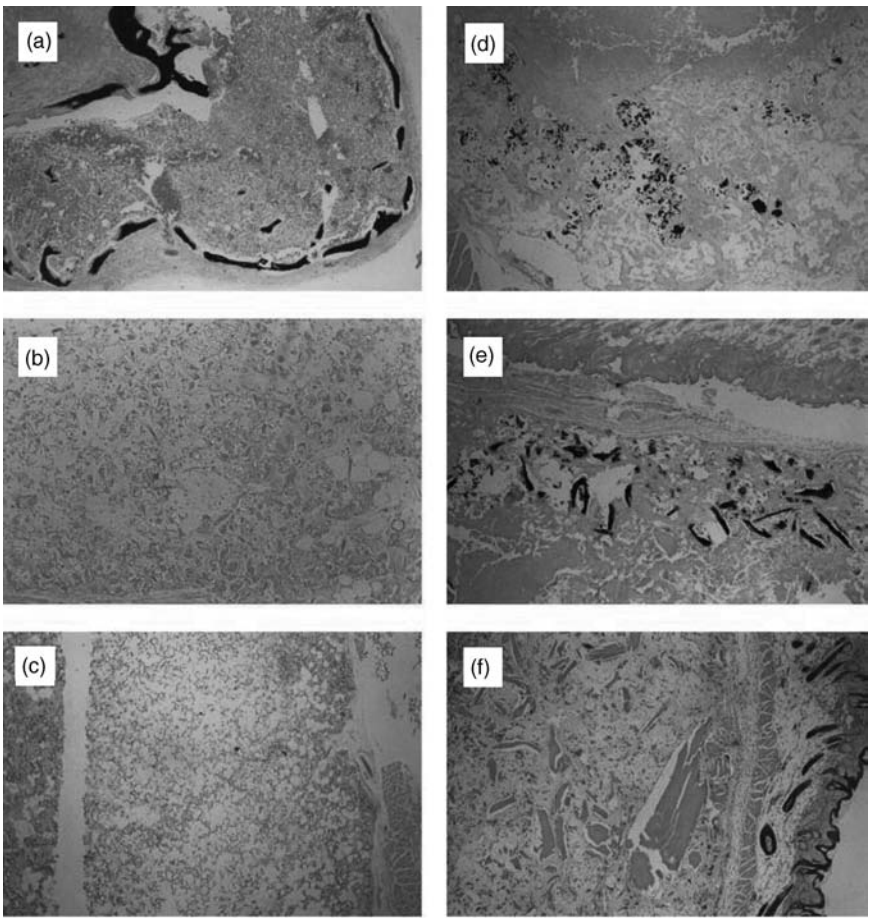


Figure 12. Photomicrographs of von Kossa and H&E histological sections of implanted (a) PGA nonwoven, (b) PLA scaffold only without GA treatment, (c) PLA scaffold only with GA treatment, (d) SIS/PLA scaffold without GA treatment, (e) SIS/PLA scaffold with GA treatment, and (f) SIS/PLA scaffold with GA treatment (H&E) (Original magnification 100×).

sometimes cause polymer deterioration even when sterilization takes place at room temperature. However, the time of exposure is relatively short (overnight), and most scaffolds can be sterilized with this method. The cold EOG sterilization method is the most widely used, with conditions of 35°C and 95% humidity. While the hot EOG method, which uses 60°C and 95% humidity, can cause shrinkage of the PLGA scaffold. One significant problem is residual EOG, which is harmful on the surface and within the polymer. Therefore, it is important that the scaffolds are subjected to adequate degassing or aeration subsequent to EOG sterilization, so that the concentration of residual EOG can be reduced to acceptable levels.

Radiation sterilization using isotopic ^{60}Co can also deteriorate polymers, since at high dosages the polymer chains can be dissociated or crosslinked according to the characteristics of the chemical structures. At a 2.5 Mrad dose, the tensile strength and molecular weight of PLGA decreases. Also, there is a rapid decrease in the molecular weight of the PGA nonwoven felt with increasing doses of radiation. It is important to remember that the properties and useful lifetime of the PLGA implant can be significantly affected by irradiation. In the case of polyethylene, it becomes a brittle and hard material at doses as high as 25 Mrad; This is due to a combination of random chain scission crosslinking. Polypropylene will often discolor during irradiation giving the product an undesirable tint, but a more severe problem is the embrittlement resulting in flange breakage, luer cracking, and tip breakage. The physical properties continued to deteriorate with time following irradiation.

Sterilization methods might significantly affect the physicochemical properties of the scaffold matrix. The specific effects with various methods are determined by the kinds of scaffold materials themselves, the scaffold preparation methods, and the sterilization factors. It is essential that a new standard for sterilizing scaffold devices be designed and established.

CONCLUSIONS

Tissue engineering, including regenerative medicine in recognition of its tremendous potential, has received a revolutionary "research push." As a result, there have been many reports on the successful regeneration of tissues and organs including skin, bone, cartilage, the peripheral and central nerves, tendon, muscle, cornea, bladder and urethra, and liver as well as composite systems like the human phalanx and joint, using scaffold biomaterials from polymers, ceramic, metal, composites and its hybrids. As previously emphasized, scaffold materials must contain a site of cellular and molecular induction and adhesion, and must allow for the migration and proliferation of cells through porosity. They must also maintain strength, flexibility, biostability, and biocompatibility to mimic a more natural, 3D environment. From this standpoint, control over a precise biochemical signal must be fostered by the combination

of a scaffold matrix and bioactive molecules including genes, peptide molecules, and cytokines. Moreover, the combination of cells and redesigned bioactive scaffolds should expand to a tissue level of hierarchy. To achieve this goal, novel scaffold biomaterials, scaffold fabrication methods, and characterization methods must be developed.

ACKNOWLEDGMENTS

This work was supported by grants from the Korea Ministry of Wealth and Health (0405-BO01-0204-0006) and stem cell Research Center (SC3100).

BIBLIOGRAPHY

Cited References

1. Langer R, Vacanti J. Tissue engineering. Science 1993;260: 920–926.
2. Nerem RM , Sambanis A. Tissue engineering: from biology to biological substitutes. Tissue Eng 1995;1:3–13.
3. Griffith LG, Naughton G. Tissue engineering—Current challenges and expanding opportunity. Science 2002;295:1009–1014.
4. Baldwin SP, Saltzman WM. Materials for protein delivery in tissue engineering. Adv Drug Deliv Rev 1998;33:71–86.
5. Mann BK, West JL. Tissue engineering in the cardiovascular system: Progress toward a tissue engineered heart. Anat Record 2001;263:367–371.
6. Lee HB, Khang G, Lee JH. Chapter 3, Polymeric biomaterials. In: Park JB, Bronzino JD, editors. Biomaterials: Principles and Applications. Boca Raton: CRC Press; 2003.
7. Patrick CW, Jr. Tissue engineering strategies for adipose tissue repair. Anat Record 2001;263:361–376.
8. Petit-Zeman S. Regenerative medicine. Nature Biotech 2001;19:201–206.
9. Hench LL, Polak JM. Third-generation biomedical materials. Science 2002;295:1014–1017.
10. Seal BL, Otero TC, Panitch A. Polymeric biomaterials for tissue and organ generation. Mater Sci Eng 2001;R34:147–230.
11. Babensee JE, McIntire LV, Mikos AG. Growth factor delivery for tissue engineering. Pharm Res 2000;17:497–504.
12. Chaignaud BE, Langer R , Vacanti JP. Chapter 1, The history of tissue engineering using synthetic biodegradable polymer scaffolds and cells. In: Atala A, Mooney DJ, editors. Synthetic Biodegradable Polymer Scaffolds. Boston: Birkhauser; 1996.
13. Rose FRA, Oreffo ROC. Bone tissue engineering: Hope vs Hype. Biochem Biophys Res Commun 2002;292:1–7.
14. Freyman TM, Yannas IV, Gibson LJ. Cellular materials as porous scaffolds for tissue engineering. Prog Mater Sci 2001;46:273–282.
15. Woolverton CJ, Fulton JA, Lopina ST, Landis WJ. Chapter 3, Mimicking the natural tissue environment. In: Lewandrowski K-U, Wise DL, Trantolo DJ, Gresser JD, Yasemski MJ, Altobeli DE, editors. Tissue Engineering and Biodegradable Equivalents: Scientific and Clinical Applications. New York: Marcel Dekker; 2002.
16. Tabata Y. The importance of drug delivery systems in tissue engineering. PSTT 2000;3:80–89.
17. Wong WH, Mooney DJ. Chapter 4, Synthesis of properties of biodegradable polymers used as synthetic matrices for tissue engineering. In: Atala A, Mooney DJ, editors. Synthetic Biodegradable Polymer Scaffolds. Boston: Birkhauser; 1996.
18. Rwoley JA, Madlambayan G, Mooney DJ. Alginate hydrogels as synthetic extracellular matrix. Biomaterials 1999;20:45–53.
19. Shakibaei M, De Souza P. Differentiation of mesenchymal limb bud cells to chondrocytes in alginate bead. Cell Biol Int 1997;21:75–86.
20. Madihally SV, Matthew HW. Porous chitosan scaffolds for tissue engineering. Biomaterials 1999;20:1133–1142.
21. Kang HW, Tabata Y, Ikada Y. Fabrication of porous gelatin scaffolds for tissue engineering. Biomaterials 1999;20:1339–1344.
22. Dunn CJ, Goa KL. Fibrin sealant: A review of its use in surgery and endoscopy. Drugs 1999;58:863–886.
23. Mayne R, Burgeson RE. Structure and function of collagen types. In: Mecham RP, editor. Biology of extracellular matrix: A Series. Orlando: Academic Press; 1987.
24. Li S-T. Chapter 6, Biologic biomaterials: Tissue-derived biomaterials (Collagen). In: Park JB, Bronzino JD, editors. Biomaterials: Principles and Applications. Boca Raton (FL): CRC Press; 2003.
25. Schense JC, Bloch J, Aebischer P, Hubbell JA. Enzymatic incorporation of bioactive peptides into fibrin matrices enhances neurite extension. Nature Biotech 2000;18:415–419.
26. Zacchi V, Soranzo C, Cortivo R, Radice M, Brun P, Abatangelo G. In vitro engineering of human skin-like tissue. J Biomed Mater Res 1998;40:187–194.
27. Malette WG, Quigley HJ, Gaines RD, Johnson ND, Rainer WG. Chitosan: a new hemostatic. Ann Thorac Surg 1983;36: 55–58.
28. Sechriest VF, Miao YJ, Niyibizi C, Westerhauzen-Larson A, Matthew HW, Evans CF, Fu FH, Suh J-K. GAG-augmented polysaccharide hydrogel: a novel biocompatible and biodegradable material to support chondrogenesis. J Biomed Mater Res 2000;49:534–541.
29. Lee YM, Park YJ, Lee SJ, Ku Y, Han SB, Choi SM, Klokkevoid PR, Chung CP. Tissue engineered bone formation using chitosan/tricalcium phosphate sponges. J Periodontol 2000; 71:410–417.
30. Lee DA, Noguchi T, Knight MM, O'Donnell L, Bently G, Bader DL. Response of chondrocyte subpopulations cultured within unloaded and loaded agarose. J Orthop Res 1998;16: 726–733.
31. Lee DA, Frean SP, Lee P, Bader DL. Dynamic mechanical compression influences nitric oxide production by articular chondrocytes seeded in agarose. Biochem Biophys Res Commun 1998;251:580–585.
32. Badylak SF, Record R, Lindberg K, Hodde J, Park K. Small intestine submucosa: a substrate for in vitro cell growth. J Biomater Sci, Polymer Ed 1998;9:863–878.
33. Khang G, Shin P, Kim I, Lee B, Lee SJ, Lee YM, Lee HB, Lee I. Preparation and characterization of small intestine submucosa particle impregnated PLA scaffolds: The application of tissue engineered bone and cartilage. Macromol Res 2002;10:158–167.
34. Badylak SF. The extracellular matrix as a scaffolds for tissue reconstruction. Cell Develop Biol 2002;13:377–383.
35. Gustafson C-J, Katz G. Cultured autologous keratinocytes on a cell-free dermis in the treatment of full-thickness wounds. Burns 1999;25:331–335.
36. Williams SF, Martin DP, Horowitz DM, Peoples OP. PHA applications: Addressing the price performance issue, I. Tissue Engineering. Int J Biolog Macromol 1999;25:111–121.
37. Khang G, Lee HB. Chapter 67. Cell-synthetic surface interaction: Physicochemical surface modifications. In: Atala A, Lanza R, editors. Orlando: Academic Press; 2001.
38. Khang G, Seong H, Lee HB. Sustained delivery of drugs with biodegradable. In: Hsuie GH, Okano T, Kim YU, Sung W-W, Yui N, Park KD, editors. Taipei, Taiwan: Princeton International Publishing Co.; 2002.
39. Khang G, Lee SJ, Han CW, Rhee JM, Lee HB. Preparation and characterization of natural/synthetic hybrid scaffolds.

In: Elcin M, editor. London, England: Kluwer-Plenum Press; 2003.
40. Khang G, Lee JH, Lee I, Rhee JM, Lee HB. Interaction of different types of cells on PLGA surfaces with wettability chemogradient. Macromol Res 2000;8:276–284.
41. Khang G, Choi MK, Rhee JM, Rhee SJ, Lee HB, Iwasaki Y, Nakabayashi N, Ishihara K. Biocompatibility of poly(MPC-*co*-EHMA)/PLGA blends. Macromol Res 2001;9:107–115.
42. Khang G, Park CS, Rhee JM, Lee SJ, Lee YM, Lee I, Choi MK, Lee HB. Preparation and characterization of demineralized bone particle impregnated PLA scaffolds. Macromol Res 2001;9:267–276.
43. Choi HS, Khang G, Shin H-C, Rhee JM, Lee HB. Preparation and characterization of fentanyl-loaded PLGA microspheres; *In vitro* release profiles. Int J Pharm 2002;234:195–203.
44. Lee SJ, Khang G, Lee YM, Lee HB. Interaction of human chondrocyte and fibroblast cell onto chloric acid treated poly(α-hydroxy acid) surface. J Biomater Sci, Polym Ed 2002;13:197–212.
45. Khang G, Choi CW, Rhee JM, Lee HB. Interaction of different types of cells on physicochemically treated PLGA surfaces. J Appl Polym Sci 2002;85:1253–1262.
46. Khang G, Jeon EK, Rhee JM, Lee I, Lee SJ, Lee HB. Controlled release of NGF from sandwiched PLGA films for the application of neural tissue engineering. Macromol Res 2003; 11:334–340.
47. Jang JW, Lee B, Han CW, Lee I, Lee HB, Khang G. Preparation and characterization of ipriflavone-loaded PLGA scaffolds for tissue engineered bone. Polymer(Korea) 2003;27:226–234.
48. Khon J, Langer R. Chapter 2.5, Bioresorbable and bioerodible materials. In: Ratner BD, Hoffman AS, Scheon FJ, Lemons JE, editors. Biomaterials Science: An Introduction to Materials in Medicine, San Diego: Academic Press; 1996.
49. Vacanti CA, Langer R, Schloo B, Vacanti JP. Synthetic polymers seeded with chondrocytes provide a template for new cartilage formation. Plast Reconstr Surg 1991;88:753–759.
50. Burg KJL, Porter S, Kellam JF. Biomaterials developments for tissue engineering. Biomaterials 2000;21:2347–2359.
51. Gutowska A, Jeong B, Jasionowski M. Injectable gel for tissue engineering. Anat Record 2001;263:342–349.
52. Suggs LJ, Krishna RS, Garcia CA, Peter SJ, Anderson JM, Mikos AG. In vitro and in vivo degradation of poly(propylene fumarate-*co*-ethylene glycol) hydrogel. J Biomed Mater Res 1998;42:312–320.
53. Harris JM, editor. Poly(ethylene glycol) Chemistry: Biotechnical and Biomedical Applications. New York: Plenum Publish. Co.; 1997.
54. Qui Y, Park K. Environment-sensitive hydrogels for drug delivery. Adv Drug Deliv Rev 2001;53:321–339.
55. Webb D, An YH, Gutowska A, Mironov VA, Friedman RJ. Propagation of chondrocytes using thermosensitive polymer gel culture. Orthoped J Musc Orthoped Surg 2000;3:18–22.
56. Sims CD, Butler P, Casanova R, Lee BT, Randolph MA, Lee A, Vacanti CA, Yaremchuk MJ. Injectable cartilage using polyethylene oxide polymer substrate. Plast Reconstruct Surg 1996;95:843–850.
57. Laurencin CT, El-Amin SF, Ibim SE, Willoughby DA, Attawia M, Allcock HR, Ambrosio AA. A highly porous 3-dimensional polyphosphazene polymer matrix for skeletal tissue engineering. J Biomed Mater Res 1996;30:133–138.
58. Agarwal S, Gassner R, Piesco NP, Ganta SR. Chapter 7, Biodegradable urethanes for biomedical applications. In: Lewandrowski K-U, Wise DL, Trantolo DJ, Gresser JD, Yasemski MJ, Altobeli DE, editors. Tissue Engineering and Biodegradable Equivalents: Scientific and Clinical Applications. New York: Marcel Dekker; 2002.
59. Zhang JY, Beckman EJ, Piesco NP, Agarwal S. A new peptide based urethane polymer: synthesis, degradation, and potential to support cell growth in vitro. Biomaterials 2000;21:1247–1258.
60. Billotte WG. Chapt. 2, Ceramic biomaterials. In: Park JB, Bronzino JD, editors. Biomaterials: Principles and Applications. Boca Raton (FL): CRC Press; 2003.
61. Hench LL. Bioactive ceramics. Ann NY Acad Sci 1988; 523:54–71.
62. Frician JC, Bareille R, Rouais F. In vitro dissolution of coral in periodontal or fibroblast cell culture. J Dent Res 1998;77:406–411.
63. Yoshikawa T, Oghushi H, Uemura T. Human marrow cells derived cultured bone in porous ceramics. Bio-Med Mater Eng 1998;8:311–320.
64. Unpublished data.
65. Krewson C, Dause R, Mak M, Saltzman WM. Stabilization of nerve growth factor in controlled release polymers and in tissue. J Biomater Sci, Polym Ed 1996;8:103–117.
66. Haller MF, Saltzman WM. Localized delivery of proteins in the brain. Pharm Res 1998;15:377–385.
67. Ito Y, Lui SQ, Imanishi Y. Enhancement of cell growth on growth factor-immobilized polymer films. Biomaterials 1991; 12:449–453.
68. Khul PR, Grriffith-Cima LG. Tethered epidermal growth factor as a paradigm for growth factor-induced stimulation from the solid phase. Nature Med 1996;2:1022–1027.
69. Duncan R, Spreafico F. Polymer conjugates. Pharmacokinetic considerations for design and development. Clin Pharmacokinet 1994;27:290–306.
70. Massia SP, Hubbell JA. Covalent surface immobilization of Arg-Gly-Asp- and Tyr-Ile-Gly-Ser-Arg-containing peptides to obtain well-defined cell-adhesive substrate. Anal Biochem 1990;187:292–301.
71. Leibmann-Vinson A, Hemperly JJ, Guarino RD, Spargo CA, Heidaran MA. Chapter 36, Bioactive extracellular matrices: Biological and biochemical evaluation. In: Lewandrowski KU, Wise DL, Trantolo DJ, Gresser JD, Yasemski MJ, Altobeli DE, editors. Tissue Engineering and Biodegradable Equivalents: Scientific and Clinical Applications. New York: Marcel Dekker; 2002.
72. Thompson RC, Wake MC, Yasemski MJ, Mikos AG. Biodegradable polymer scaffolds to regenerate organs. Adv Polym Sci 1995;122:245–274.
73. Khang G, Jeon JH, Cho JC, Lee HB. Fabrication of tubular porous PLGA scaffolds by emulsion freeze drying methods. Polymer(Korea) 1999;23:471–177.
74. Bowlin GL, Pawlowski KJ, Boland ED, Simpson DG, Fenn JB, Wnek GE, Stitzel JD. Chapter 9, Electrospinning of polymer scaffolds for tissue engineering. In: Lewandrowski K-U, Wise DL, Trantolo DJ, Gresser JD, Yasemski MJ, Altobeli DE, editors. Tissue Engineering and Biodegradable Equivalents: Scientific and Clinical Applications. New York: Marcel Dekker; 2002.
75. Athanasious KA, Neiderauer GG, Agrawal CM. Sterilization, toxicity, biocompatibility and clinical applications of polylactic acid/polyglycolic acid copolymers. Biomaterials 1996;17:93–102.

Reading List

Jeon EK, Khang G, Lee I, Rhee JM, Lee HB. Preparation and release profile of NGF-loaded PLA scaffolds for tissue engineered nerve regeneration. Polymer(Korea) 2001;25:893–901.

See also ENGINEERED TISSUE; STERILIZATION OF BIOLOGIC SCAFFOLD MATERIALS.

BIOMECHANICS OF EXERCISE FITNESS

GIDEON ARIEL
Ariel Dynamics
Canyon, California

INTRODUCTION

Normal human development spans a lifetime from infancy to old age. Modern civilization is confronted with the lengthening of that time and its effect on the individual and society. Housing improvements, employment alterations, labor saving devices, and modern medicine are but a few of the factors protecting humanity from those instances which previously shortened life. While many of the difficult, threatening experiences have been eliminated or reduced in severity, problems remain to be solved. Concerns for the quality of life as people become older include maintaining self-sufficiency. Many solutions conflict with beliefs generally termed "current wisdom" in areas, such as training, dieting, exercising, and aging. While society ages, the challenge for each individual is to strive to retain the lowest "biological" age while their "chronological" birthdays increase. The dilemma concerns the best way to accomplish this task.

The main purpose of this article is to focus on the biomechanical principles of movement, the scientific bases of training and fitness, and the optimization of human performance at any age. These are not just nonsense concepts added to the quantities of known theories, but are objectively quantifiable procedures that encompass our understandings and can produce precise conclusions. Mathematical principles and gravitational formulations provide the cornerstones for optimizing human performance. Biological, anatomical, physiological, and medical discoveries are always under investigation, challenge, and improvement and these findings will be incorporated into many of the current theories. Figure 1 illustrates just part of the anatomy and its complicated structure. The struggle will continue among scientists to establish new principles for revolutionizing the world of gerontology, diet, physical fitness and training, and amplifying those factors necessary for extending life not only in length, but also in quality. Scientists with expertise in many different areas will be addressing the problems associated with aging from their specialized perspective.

In order to address the optimization of human movement and performance, the underlying philosophical premise metaphorically compares life with sport. The goal is that everyone should be a gold medalist in their own body regardless of age. Most people, however, do not achieve their Gold Medal because their goals, potential, and/or timing are uncoordinated or nonexistent. For example, an individual may envision themselves as a tennis champion, yet lack the requisite physical and physiological traits of the greatest players. Given this situation, can a person's potential be maximized? Achieving one's maximum potential necessitates tools applicable to everyone for improving their performance, whether in tennis, fitness, overcoming physical handicaps, or fighting disease. Useful tools must be based, however, on correct, substantive scientific principles.

SCIENTIFIC PRINCIPLES FOR QUANTIFYING MOTION

Human movement has fascinated humans for centuries including some of the world's greatest thinkers, such as Leonardo da Vinci, Giovanni Borelli, Wilhelm Braune, and others. Many questions posed by these stellar geniuses have been or can be addressed by the relatively new area of Biomechanics. Biomechanics is the study of the motion of living things, primarily, and it has evolved from a fusion of the classic disciplines of anatomy, physiology, physics, and engineering. Bio refers to the biological portion, incorporating muscles, tendons, nerves, and so on, while mechanics is associated with the engineering concepts based upon the laws described by Sir Isaac Newton. Human bodies consist of a set of levers that are powered by muscles. Quantification of movements, whether human, animal, or inanimate objects, can be treated within biomechanics according to Newtonian equations. It may seem obvious, with the perfect vision of hind sight, that humans and their activities, such as the wielding of tools (e.g., hammer, axe) or implements (e.g., baseball bat, golf club, discus), must obey the constraints of gravitational bodies, just as bridges, buildings, and cars do. For some inexplicable reason, humans and their activities had not been subjected to the appropriate engineering concepts that architects would use when determining the weight of books to be housed in a new library or engineers would apply to designing a bridge to span a wide, yawning abyss. It was not until Newton's

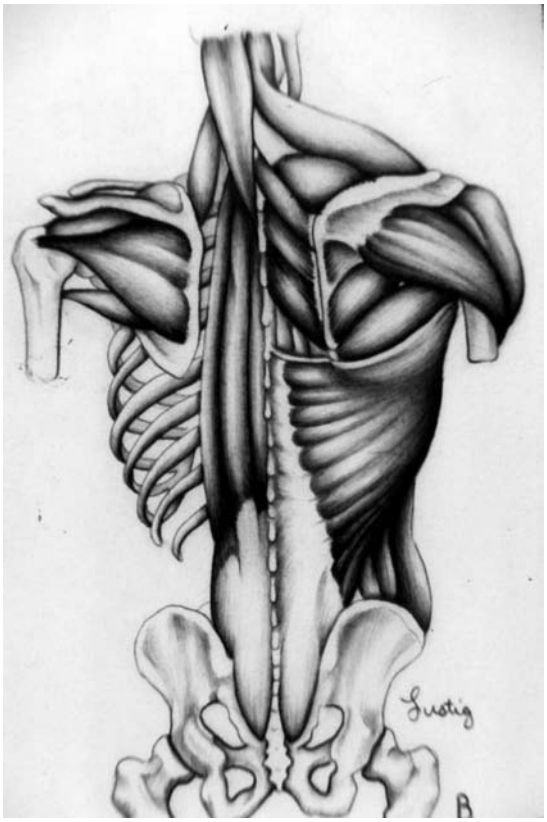


Figure 1. The human structure.

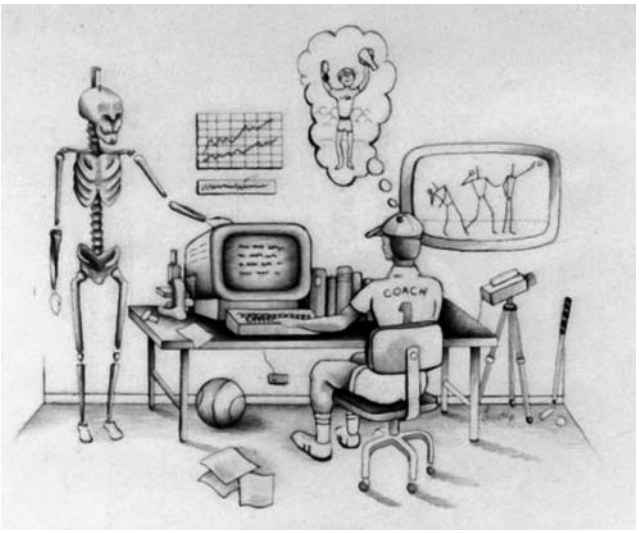


Figure 2. The modern coach and his tools.

apple fell again during the twentieth century that biomechanics was born.

Biomechanics, then, is built on a foundation of knowledge and the application of the basic physical laws of gravitational effects as well as those of anatomy, chemistry, physiology, and other human sciences. Early quantification efforts of human movement organized the body as a system of mechanical links. Activities were recorded on movie film that normally consisted of hundreds of frames for each of the desired movement segment. Since each frame of the activity had to be processed individually, the task was excessively lengthy, tedious, and time intensive. Figure 2 illustrates an abstract of todays sophisticated coaching tools in athletics. The hand calculations of a typical 16 segment biomechanical human required many hours for each frame, necessitating either numerous assistants or an individual investigator's labor-of-love and, frequently, both. Unfortunately, these calculations were susceptible to numerical errors.

The introduction of large, main-frame computers improved reliability and reasonableness of the results, replacing much of the skepticism or distrust associated with the manually computed findings. Computerization accelerated the calculations of a total movement much more rapidly than had been previously possible, but presented new difficulties to overcome. Many of the early biomechanical programs were cumbersome, time intensive main-frame endeavors with little appeal except to the obsessed, devotee of computers, and movement assessment. However, even these obstacles were conquered in the ever expanding computerization era. The computerized hardware/software system provides a means to objectively quantify the dynamic components of movement in humans regardless of the nature of the event. Athletic events, gait analyses, job-related actions as well as motion by inanimate objects, including machine parts, air bags, and auto crash dummies are all reasonable analytic candidates. Objectivity replaces mere observation and supposition.

One of the most important aspects included in the Bio portion of biomechanics is the musculoskeletal system. Voluntary human movement is caused by muscular contractions that move bones connected at joints. The neuromuscular system functions as a hierarchical system with autonomic and basic, life sustaining operations, such as heart rate and digestion, controlled at the lowest, noncognitive levels and with increasing complexities and regulatory operations, such as combing the hair or kicking a ball, controlled by centers that are further up the nervous system. Interaction of the various control centers is regulated through two fundamental techniques each governed like a servosystem.

The first technique equips each level of decision making with subprocessors that accept the commands from higher levels as well as accounting for the inputs from local feedback and environmental information sensors. Thus, a descending pyramid of processors is defined that can accept general directives and execute them in the presence of varying loads, stresses, and other perturbations. This type of input–output control is used for multimodal processes, such as maintaining balance while walking on an uneven terrain, but would be inappropriate for executing deliberate, volitional, complex tasks like the conductor using the baton to coordinate the music of the performing musicians.

The second technique utilized by the brain to control muscular contractions applies to the operation of higher level systems that generate output strategies in relation to behavioral goals. These tasks use information from certain sensory inputs, including joint angle, muscle loading, and muscular extension or flexion that are assessed, transmitted to higher centers for computation, which then executes the set of modified neural transmissions received. Cognitive tasks requiring the type of informational input that influences actions are the ones with which humans are most familiar since job execution requires more thought than breathing or standing upright. A frequently misunderstood concept is that limb movement is possible only through contractions of individual muscle fibers. For most cases of voluntary activity, muscles work in opposing pairs with one set of muscles opening or extending the joint (extensors) while the opposite muscle group closes or flexes the joint. The degree of contraction is proportional to the frequency of signals from the nerve as signaled from the higher centers. Movement control is provided by a programmable mechanism so that when flexors contract, the extensors relax, and vice versa. The motor integration programmed generated in the higher, cognitive levels regulates not only the control of the muscle groups around a joint, but also those necessary actions by other muscles and limbs to redistribute weight, to counteract shifts in the center of gravity.

One of the most important, but frequently misunderstood, concepts of the nervous system is the control and regulation of coordinated movement. When a decision is made to move a body segment, the prime muscles or agonists receive a signal to contract. The electrical burst stimulates the agonist muscular activity causing an acceleration of the segment in the desired direction. At the same time, a smaller signal is transmitted to the opposite muscle group, or antagonist, which causes it to function as a joint stabilizer. With extremely rapid movements, the antagonist is frequently stimulated to slow the limb in time to

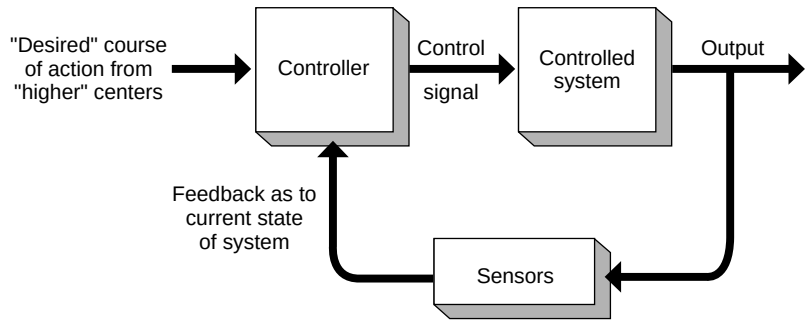


Figure 3. Feedback control mechanism of movement.

protect the joint from injury. It is the strength and duration of the electrical signal to both the agonist and antagonist that govern the desired action. The movement of agonists and antagonists, whether a cognitive process, such as throwing a ball, or an acquired activity, such as postural control, is controlled by the nervous system. Figure 3 illustrates a flowchart for the control system for movement. Many ordinary voluntary human activities resulting from agonist–antagonist muscular contraction are classified by different terms, isotonic, variable resistance, and ballistic. Slower movements, demonstrating smaller, more frequent, electrical signal alterations, are intricately controlled by both agonist and antagonist. These types of motion are tracking movements.

One control mechanism available involves the process of information channeled between the environment and the musculature. Closed-loop control involves the use of feedback whereby differences between actual and desired posture are detected and subsequently corrected, whereas open-loop control utilizes feed forward strategies that involve the generation of a command based on prior experience rather than on feedback. Braitenberg and Onesto (ARBIB) proposed a network for converting space into time by providing that the position of an input would determine the time of the output. This open loop system would trigger a preset signal from the nervous system to the muscle generating a known activity. Kicking a ball, walking, throwing a baseball, swinging a golf club, and hand writing are considered ballistic movements.

When a limb moves, a sophisticated chain of events occurs before, during, and after the movement is completed. The fineness of control depends on the number of muscle fibers innervated by each motor neuron. A motor unit is generally defined as a single motor neuron and the number of muscle fibers it innervates. Fine control is achieved when a single motor neuron innervates just a few fibers. Less fine control, as in many large muscle groups, is attained when individual motor units innervate hundreds or even thousands of fibers. The more neurons there are, the finer the ability to maneuver, as with eye movements or delicate hand manipulations. In contrast to the high innervations ratio of the eye, the biceps of the arm has a very low rate of nerve-to-muscle fiber resulting in correspondingly more coarse movements.

While the amount of nervous innervations is important when anticipating the precision of control, the manner of interaction and timing between muscles, nerves, and desired outcome is probably more important when evaluating performance. Recognizable actions elicit execution of patterned, synchronous nervous activity. Frequently repeated movements are usually performed crudely in the beginning stages of learning, but become increasingly more skilled with use and/or practice. Consider the common activity of handwriting and the execution of one's own signature. The evolution from a child's irregular, crude printing to an adult's recognizable, consistently repeatable signature is normal. Eventually, the individual's signature begins to appear essentially the same every time and is uniquely different from any other person. Not only can the person execute handwritten signatures consistency, but can use chalk to sign the name in large letters on a blackboard producing a recognizably similar appearance. The individuality of the signature remains whether using the fine control of the hand or recruiting the large shoulder and arm muscles not normally required for the task. Reproduction of recognizable movements occurs from preprogrammed control patterns stored in the brain and recruited as necessary. Practicing a golf swing until it results in a 300 yard drive down the middle of the fairway, getting the food-laden fork from the plate into the mouth, and remembering how to ride a bike after a 30 year hiatus illustrate learned behavior that has become "automatic" with practice and can be recalled from the brain's storage for execution.

Volitional tasks require an integration of neurological, physiological, biochemical, and mechanical components. There are many options available when performing a task, such as walking, but eventually, each person will develop a pattern that will be recognizable as that skill, repeatable, and with a certain uniqueness associated with that particular individual. Although any person's movement could be quantified with biomechanical applications and compared to other performers in a similar group, for example, the gold, silver, and bronze medalists in an Olympic event, perhaps it will be the ability to compare one person to themselves that will provide the most meaningful assistance in the assault on aging.

There are many areas of daily living in which biomechanical analyses could be useful. Biomechanics could be utilized to design a house or chair to suit the body or to lift bigger, heavier objects with less strain. This science could be useful in selecting the most appropriate athletic event for children or for improving an adult's performance. With increasing international interest in competitive athletics, it was inevitable that computers would be used for the analysis of sports techniques. Computer calculations can provide information that surpasses the limits of what the

human eye can see and intuition can deduce. Human judgment, however, is still critically important. As in business and industry, where decisions are based ultimately upon an executive's experience and interpretive ability, the coach or trainer is, and will remain, the ultimate decision maker in athletic training. Rehabilitation and orthopedic specialists can assess impaired movement relative to normal performance and/or apply computerized biomechanical techniques to the possibilities of achieving the restoration of normal activities. With the increase in the population of older citizens, erotological applications will increase. The computer should be regarded as one more tool, however, complex, which can be skillfully used by humans in order to achieve a desired end.

One factor that humans have lived with is change. The environment in which we live is changing during every one of the $\sim$ 35 million min of our lives. The human body itself changes from birth to maturity and from maturity to death. The moment humans first picked up a stone to use as a tool, the balance between humans and the environment was altered. After that adaptation, the ways in which the surrounding world changed resulted in different effects and these were no longer regular or predictable. New objects were created from things that otherwise would have been discounted. These changes were made possible by humans due to the invention of tools. The more tools humans created, the faster was the rate of environmental change. The rate of change due to tools has reached such a magnitude that there is danger to the whole environment and frequently to the people who use the tools, such as occurred during the Industrial Revolution, as well as in our own times with such problems as carpal tunnel syndrome. Human beings seem to have become so infatuated with their ability to invent things that they have concentrated almost exclusively upon improving the efficiency, safety, durability, cost, or aesthetic appeal of the device. It is ironic that with all of the innovative development, little consideration has been given to the most complex system with the most sophisticated computer in the world: the human body.

When they talk about their physical goals in work or in sports, people usually say they would like to do their best, meaning, reach their maximum output. It is a matter of achieving their absolute limit in speed, strength, endurance or skill and combining the elements with accuracy. This is no different than an athlete training for maximum performance in the Olympic games. The difficulty with focusing everything on maximum performance is that only a single goal, getting the highest results—fastest, biggest, quickest, longest, or most graceful—is considered a superlative or acceptable achievement. Maximums do not take into consideration other aspects of body performance that often prove to be just as important to the individual. Emphasis upon the demands for maximum performance is frequently portrayed with the thought that Winning isn't everything, it's the only thing. Figure 4 illustrates todays sophisticated biomechanical system to quantify human performance.

Imagine for a moment a maximum performance in the car industry—the perfect automobile. It is incredibly graceful and the aerodynamic, functional lines make it a thing of beauty. It accelerates from 0 to 60 miles $\cdot$ h^{-1} within a few seconds. It brakes, corners, and steers with a fineness that would permit a shortsighted 75-year old to compete at Le Mans. The suspension is so smooth that a passenger can pour liquids without spilling a drop. The car requires only minimal maintenance while averaging 50 miles $\cdot$ gal^{-1} in city driving. Best of all, it is the vehicle of the common man at a price of $5000. If all that sounds impossible—it is. Incorporating all of these maximums into a single automobile exceeds the ability of any designer or manufacturer. Instead, the individual shopping for a car must choose the attributes he or she feels are most important.

Figure 4. The modern biomechanical system.

Therein lies the problem, some goals are partly, if not wholly, incompatible with others. An automatic transmission uses more gas than a standard shift, but it does make driving easier. Sleek aerodynamic lines add grace and reduce drag, but they can also lessen head room. High performance engines provide power, but require constant care. The solution is a compromise, a willingness to make tradeoffs.

This same spirit of compromise, of accepting something less than a single maximum, should govern the operation of the most important machine in our lives—our body. Reality must be applied when comparing ourselves to Olympic athletes or, with the progression of age, mimicking various youthful physical activities. For example, there is no need to have an endurance capacity equal to the current gold medalist or the strength level equivalent to the World heavyweight record holder. Likewise, senior citizens may resist relinquishing their drivers' licenses despite their slower reaction times, poorer eyesight, and/or hearing, as well as frequently suffering from some type of chronic disease that may further reduce their strength, joint mobility, or even cognitive processes, such as memory or decision making.

Instead of a maximum, what most people really want from their bodies is to optimize their performances and lives. They seek the most efficient use of energy, of bodily action consonant with productive output, health, and enjoyment. Many people are beginning to appreciate that certain types of exercise add to the vitality of the

cardiovascular system, lessen the risk of heart attack, and make it possible to live longer and more active lives. In other words, the willingness to sacrifice 20 yards on a drive off the golf tee may mean that the golfer's feet will be able to walk the entire course without being tortured during every step. The desire is to play a couple of hours of winning tennis, stroking the ball with pace and purpose, but not if the extra zing means a tennis elbow that will be sore for several weeks. Sensible joggers prefer to run 6 rather than 10 miles a day in 40 min, if the latter leads to tender knees and shin splints. In other words, human beings must compromise between anatomy (the structural components) and physiology (the bodily processes). A correct balance between the two, at all ages, will assist in optimizing bodily efficiency.

In addition to the desire for our internal environment to be physical fit, pertinent questions should be posed about our external environment. For example, is it really necessary for that designer chair to cause a bone ache deep in the buttocks after sitting for 5 min? Can a person not spend a day laboring over a desk or piece of machinery without feeling as if a rope had been tightly tied around the shoulders at the end of the project? Why must a weekend with shovel or rake inevitably produce lower back pain on Monday? Why is it that some individuals who are 50 years old seem able to work and play as if 10–20 years younger, while some 30 year olds act as if infected with a malignant decrepitude? The answer is that, as with the anatomy and physiology achieving optimal coordination, so should the whole human organism coordinate better with its environment.

Perhaps these examples could be dismissed as the minor aches of a hypochondriac society overly concerned with its comfort. But the overall health facts for the United States and many other modern civilizations appall even those jaded by constant warnings of disaster. The American Heart Association, in urging the 2005 Congress to fund prevention programs, contends that the Number One killer of Americans is heart disease, stroke, and other cardiovascular diseases. In addition, a total of 75 million Americans are afflicted with chronic disease. On any given day, >1 million workers do not show up for their jobs because of illness, and sickness prevents a million of these from returning in <1 week. Twenty-eight million Americans have some degree of disability. Perhaps not coincidentally, a quarter of the population is classified as overweight. At least 3 million citizens have diabetes, and one-half are unaware of the problem, and the United States accounts for most of the deaths due to cardiovascular disease. The health profile of the future, the condition of the youth of today, offers no comfort. About 1 in 5 youngsters still cannot pass even a simple test of physical performance. More than 9 million American children under the age of 15 have a chronic ailment. From one-third to one-half of U.S. children are overweight and one-third of America's young men fail to meet military physical fitness requirements.

In pursuit of technological achievement, Americans have almost ignored the one major element besides food and rest needed to sustain the human body: physical activity. This has lent impetus to a subtle yet deadly disease that has reached epidemic proportions in this country and others. Cardiovascular disease is often referred to as hypokinetic disease or lack of-motion disease. Unfortunately, degeneration with Americans begins earlier rather than later. One study indicates that middle age characteristics start to show at approximately age 26. The peak age for heart disease among American men is 42 years. In Europe, it is 10 years later. A corporate wide employee health survey conducted by a large computer manufacturer indicated that smokers have 25% higher healthcare costs and 114% longer hospital stays than nonsmokers. People who did not exercise have 36% higher healthcare costs and 54% longer hospital stays than people who did exercise. Overweight people have 7% higher healthcare costs and 85% longer hospital stays than people who are not. In general, people with poor health habits have higher healthcare costs, longer hospital stays, lower productivity, more absenteeism, and more chronic health problems than those who do not. Some questions both workers and their companies should ask are (*1*) How many heart attacks, strokes, cancers, or coronary by-pass operations did your company pay for last year? (*2*) How much better would profits have been if heart diseases had been reduced 10, 20, or 30%? (*3*) How much would corporate profits increase if employee healthcare costs were reduced by 10%?

One large U.S. corporation developed a comprehensive wellness program at numerous sites. During the first year, grievances decreased by 50%, on-the-job accidents by 50%, lost time by 40%, and sickness and accident payments by 60%. The corporation estimated at least a 3:1 return per dollar invested.

The requirement for such an optimum way of life is a scientific analysis of the way people live and use their bodies. Only after such a quantitative examination can a concept of cost be determined or a better way of doing something that is more efficient and less damaging to the body, discovered. For example, rapid weight loss may result from running long distances, such as 15 miles a day, fasting drastically, or performing aerobics for 5 h a day. However, such excessive training regimens may be as detrimental to the body as sitting all day in an easy chair and simply ignoring one's obesity.

Evolution, culture, and the changing demands of existence have tended to develop forces and stresses upon the body that are not necessarily in harmony with the basic design and structure of the human equipment. Standing upright, humans employ one pair of extremities for support and the other pair capable of tremendous versatility. It would seem that of all animals, humans, fortuitously assisted by the evolution of their brain and other organs, optimized the use of their body. Unfortunately, the human body has had to pay a stiff price for its upright posture. Human vertical posture is inherently unstable; therefore, humans must devote more neuromuscular effort and control to maintain balance, than four-legged animals. There is a tendency to lean forward, which adds to the ability to move in that direction, but increases the risk of falling.

A complex neuromuscular process is constantly at work to prevent humans from toppling. Many things may interfere with this balancing act, such as consuming too much whiskey or walking on an icy sidewalk. These interruptions

of the flow of information to and from the brain centre which coordinates the balancing process can result in staggering or falling. This postural condition creates a constant strain on all the muscles employed to retain balance and upon the set of bones forming the spine. The spine is basically a tower of I-beams supports the skeletal frame and, in order to remain in good health, proper mechanical alignment is essential. Any deviation from this mechanical alignment will result in pain relating to non-alignment, such as low back or neck pain. The vulnerability of the back is threatened frequently by work, recreation situations, and furnishings, since their uses subject an already tenuous upright position to undergo increased stresses. As the body compensates for alignment problems by creating excess bone tissue and neural pain, certain arthritic conditions may be the result.

Correction or prevention in tools or activities may assist in the optimization of performance and in more closely aligning the biological with the chronological age. Clearly, optimization and compensation may conflict within the human mechanism since a logical idea may violate physical principles. Based on this introduction of merely a few of the internal and external challenges to the human organism, the need for adequate and accurate assessments, improved tools, and human behavioral modifications becomes more apparent.

With each passing year, the composition of the population in America and probably many other modern societies is becoming older. This population increase of older citizens appears to be due, in part, to the large number of individuals of all ages who are experiencing modifications of lifestyle in a variety of ways, including better working conditions, improved health–medical opportunities, and changing activity levels. Pollock et al. (1) noted that the activity levels of elderly people have increased during the previous 20 years. However, it was estimated that only 10% of elderly individuals participate in regular vigorous physical activity and that 50% of the population who are 60 or more years of age described their lifestyles as sedentary.

Scientific studies and personal experiences continue to link many of the health problems and physical limitations found in the aged to lifestyle. Sedentary living appears to be a major contributor to the significantly adverse effect on health and physical well being. Certainly, there is increasing evidence indicating the vital need for improved national and international policies for better fitness, health, and sports for older individuals. In order to address some of these indicators, new attitudes and policies must emphasize activities and resources to meet the minimal requirements for keeping older people in good health, preventing their deterioration with age, and meeting the special interests of individuals with various disorders. In addition to the difficulties that hospitals, insurance companies, children of the elderly, and legislators face, the medical and scientific communities require time to determine the most appropriate solutions for improving the quality of these lengthening lives.

Many of the myths about aging are being disproved while the true nature of age-related changes appears to be less bleak than previously thought. Disuse and disease, not age alone, are increasingly, revealed as culprits. There is an increasing awareness of the need for more emphasis on fitness to maintain wellness and prevent degenerative illness, for more research to understand the aging body of the healthy older person, and to determine the exercise needs of the ill and/or the handicapped. Pollock et al. (1) noted that physical capacity decrements are normally associated with the aging process. This loss has been attributed to the influence of disease, medication, age, and/or sedentary lifestyle. Additionally, it was noted that the majority of the elderly do not exercise and that it is unclear whether the reduced state of physical conditioning associated with aging results from the deconditioning due to sedentary living, age, or both.

It is a fact of life that muscle tissue suffers some diminution from age. Age-associated changes in organ and tissue function, such as a decline in fat-free mass, total body and intracellular water, and an increase in fat mass (2) may alter the physiological responses to exercise or influence the effect(s) of medication. However, any discussion about age realistically utilizes arbitrary time periods apportioned eons ago by men who evaluated time relative to the number of revolutions of the earth around the sun and the rotation of the earth on its own axis. These predetermined periods may or may not have any relationship with the aging of the cells in the body. The linkage between the chronological age and the biological age of people is imprecise. Perhaps a more accurate consideration of the relationship between chronological and biological age would be one that is nonlinear, may differ with gender, or be dependent on other factors.

It is an inevitable evolutionary consequence that individuals within a species differ in many ways. The characterization of an individual on the basis of a chronological age scale may be practical, but biologically inappropriate. It may be that use or functional activities may have a greater influence on determining biological age rather than the number of times the earth has revolved around the sun. It appears that biological age can be affected by genetic code, nutrition and, most physical activity. Astrand (3) suggested that as an individual ages, the genetic code may have more of an effect on the function of systems with key importance in physical performance. He also noted that a change in lifestyle, at almost any chronological age, can definitely modify the biological age, either upward or downward. It has been suggested that the disparity of older persons is a hallmark of aging itself (4). It is important to determine how much age variance is due to the passage of time and how much is caused by the accumulation of other, nontime dependent, alterations. Previous attitudes towards physical adversities observed in the elderly were that they were attributable to disease. More recently, a third dimension associated with poor health in older persons has been described by Bortz and Bortz (4) as The Disuse Syndrome. For example, one of the most common markers of aging was thought to be a decreased lean body mass. However, analysis of 70 year old weight lifters revealed no such decline. The components of the Disuse Syndrome have been similarly grouped by Kraus and Raab (5) in their book, Hypokinetic Disease, and are (*1*) cardiovascular vulnerability; (*2*) musculoskeletal fragility; (*3*) obesity; (*4*) depression; (*5*) premature aging.

Use is a universal characteristic of life. When any part of the body has little or no use, it declines structurally and

functionally. The effects of disuse can be observed on any body part, such as atrophied intestinal mucosa, when a loop is excluded from digestive functions or the lung becomes atelectatic when not aerated. A lack of adequate conditioning and physical activity causes alterations in the heart and circulatory system, as well as the lungs, blood volume, and skeletal muscle (6–9). During prolonged bed rest, blood volume is reduced, heart size decreases, myocardial mass falls, blood pressure response to exercise increases, and physical performance capability is markedly reduced. On the other hand, although acute changes within the cardiovascular system result in response to increased skeletal muscle demands during exercise, there is evidence that chronic endurance exercise produces changes in the heart and circulation that are organic adaptations to the demands of chronic exercise (10–15).

Cardiac performance undergoes direct and indirect age-associated changes. There is a reduction in contractility of the myocardium (16) and this increased stiffness impairs ventricular diastolic relaxation and increases end diastolic pressure (17). This suggests that exercise-induced increases in heart rate would be less well tolerated in older individuals than in younger populations. The decline in maximal heart rate is known and the cause is multifactorial, but is mostly related to a decrement in sympathetic nervous system response. Fifty percent of Americans who are >65 years of age have a diagnostically abnormal resting electrocardiogram (18). Another factor associated with aging is a progressive increase in rigidity of the aorta and peripheral arteries due to a loss of elastic fibers, increase in collagenous materials, and calcium deposits (19). When aortic rigidity increases, the pulse generated during systole is transmitted to the arterial tree relatively unchanged. Therefore, systolic hypertension predominates in elderly hypertensive patients.

Other bodily systems demonstrate age-related alterations. Baroreceptor sensitivity decreases with age and hypertension (20,21) such that rapid adjustment of the cerebral circulation to changes in posture may be impaired. Kidney function reveals a defect in renal concentrating ability and sluggish renal conservation of sodium intake causes elderly patients to be more susceptible to dehydration (22). Hyaline cartilage on the articulator surface of various joints shows degenerative changes and clinically represents the fundamental alteration in degenerative osteoarthritis (23). A decrease in bone mineral density (osteoporosis) can reduce body stature as well as predispose the individual to spontaneous fractures. Older women are more prone to osteoporosis than older men and this may reflect hormonal differences (23). Older persons are less tolerant of high ambient temperatures than younger people (24) due to a decrease in cardiovascular and hypothalamic function which compromises the heat dissipating mechanisms. Heat dissipation is further compromised by the decrease in fat-free mass, intracellular and total body water, and an increase in body fat.

Unfortunately, the effects of disuse on the body manifest themselves slowly since humans normally have redundant organs that can compensate for ineffectiveness or disease. In addition, humans are opaque so that disease or deterioration are externally unobservable and, thus, go unheeded (e.g., the early changes in bones due to osteoporosis are subclinical and are normally detected only after becoming so pronounced that fractures ensue). Cummings et al. (25) mentioned the difficulty of distinguishing manifestations in musculoskeletal changes due to disease related to aging. Muscle mass relative to total body mass begins decreasing in the fifth decade and becomes markedly reduced during the seventh decade of life. This change results in reduced muscular strength, endurance, size, as well as a reduction in the number of muscle fibers. Basmajian and De Luca (26) reported numerous alterations in the electrical signals associated with voluntary muscular contractions with advancing age. As yet, there are no findings published that have definitively located age-related musculoskeletal changes in either the nervous or the muscular system. The diaphragm and cardiac muscle do not seem to incur age changes. Perhaps this is due to constant use, from exercise, or possibly a genetic survival mechanism.

There is growing consensus that many illnesses are preventable by good health practices including physical exercise. Milliman and Robertson (27) reported that, of the 15,000 employees of a major computer company, the nonexercisers accounted for 30% more hospital stays than the exercisers. Lane et al. (28) reported that regular runners had only two-thirds as many physician visits as community matched controls. The beneficial effect of exercise on diabetes has long been recognized and is generally recommended as an important component in the treatment of diabetes (29). Regular endurance exercise favorably alters coronary artery disease risk factors, including hypertension, triglyceride and high density lipoprotein cholesterol concentrations, glucose tolerance, and obesity. In addition, regular exercise raises the angina threshold (30).

Jokl (31) suggested three axioms of gerontology that are affected by exercise. He contents that sustained training results in the following: (*1*) decline of physique with age; (*2*) decline of physical fitness with age; (*3*) decline of mental functions with age.

Health in older people is best measured in terms of function, mental status, mobility, continence, and a range of activities of daily living. Preventive strategies appear to be able to forestall the onset of disease. Whether exercise can prevent the development of atherosclerosis, delay the occurrence of coronary artery disease, or prevent the evolution of hypertension is at present debatable. But moderate endurance exercise significantly decreases cardiovascular mortality (32). Endurance exercise can alter the contributions of stress, sedentary lifestyle, obesity, and diabetes to the development of coronary artery disease (33).

For example, the four-time Olympic discus champion, Al Oerter, at the age of 43, focused his training to qualify for the 1980 Olympic Games that would have been his fifth consecutive Olympiad Oerter threw his longest throw [220 f (67.05 m)] but, since the United States boycotted the 1980 Moscow Olympic Games, his chance was denied. By the time of the 1984 Los Angeles Games, Oerter was 47 years old. Even at an age well beyond most Olympic competitors, he again threw his best, exceeding 240 f (73.15 m) in practice sessions. Oerter's physique and strength suggested that his biological age was less than his chronological age.

Biologically, he was probably between 25 and 30, although chronologically he was 15–20 years older. Unfortunately, in the competition that determined which athletes would represent the United States, Oerter suffered an injury that precluded him from trying to achieve an unprecedented fifth consecutive Olympic Gold medal.

PRINCIPLES FOR EXERCISE AND TRAINING

Physical fitness and exercise have become, as previously discussed, an increasing concern at nearly all levels of American society. The goal of attaining peak fitness has existed for centuries, yet two problems continue to obfuscate understanding. The ability to assess strength and/or to exercise has occupied centuries of thought and effort. For examples, Milo the Greek lifted a calf each day until the baby grew into a bull. Since this particular procedure is not commonly available, humans have attempted to provide more suitable means to determine strength levels and ways to develop and maintain conditioning. Technology for assessing human performance in exercise and fitness evaluations, in both theory and practice, exhibits two problems. First, a lack of clearly defined and commonly accepted standards results in conflicting claims and approaches to both attaining and maintaining fitness. Second, a lack of accurate tools and techniques for measuring and evaluating the effectiveness of a given device designed to diagnose present capabilities for exercising or even to determine which exercises are appropriate to provide "fitness", regardless of age or gender. Vendors and consumers of fitness technology have lacked sound scientific answers to simple questions regarding the appropriateness of exercise protocols.

Reviewing studies conducted to determine the effects of strength training on human skeletal muscle suggests many benefits with appropriate exercise. In general, strength training that uses large muscle groups in high resistance, low repetition efforts increases the maximum work output of the muscle group stressed (34). Since resistance training does not change the capacity of the specific types of skeletal muscle fibers to develop different tensions, strength is generally seen to increase with the cross-sectional area of the fiber (35). The human body can exercise by utilizing its own mass (e.g., running, climbing, sit ups). These and other forms of nonequipment based exercises can be quite useful. In addition, there are various types of exercise equipment that allow selection of a weight or resistance and then the exercise against that machine resistance is performed.

The relationship between resistance exercises and muscle strength has been known for centuries. Milo the Greek's method of lifting a calf each day until it reached its full growth probably provides the first example of progressive resistance exercises. It has been well-documented in the scientific literature that the size of skeletal muscle is affected by the amount of muscular activity performed. Increased work by a muscle can cause that muscle to undergo compensatory growth (hypertrophy), whereas disuse leads to wasting of the muscle (atrophy).

The goal of developing hypertrophy has stimulated the medical and sports professions, especially coaches and athletes, to try many combinations and techniques of muscle overload. Attempts to produce a better means of rehabilitation, an edge in sporting activities, as a countermeasure for the adverse effects of space flight, or as a means to improve or enhance bodily performances throughout a lifetime have only scratched the surface of the cellular mechanisms and physiological consequences of muscular overload.

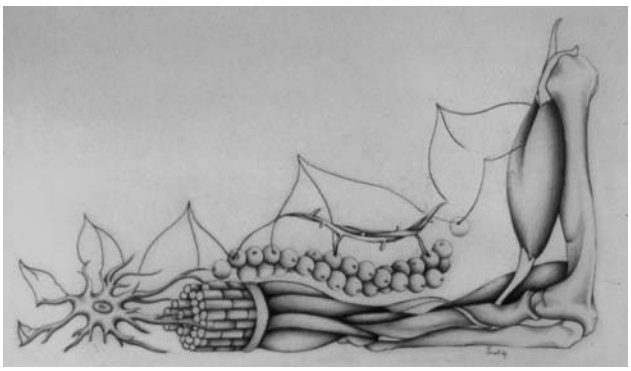


Figure 5. Integration of our muscular system.

Muscular strength can be defined as the force that a muscle group can exert against a resistance in a maximal effort. In 1948, Delorme and Watkins (36) adopted the name "progressive resistance exercise" for his method of developing muscular strength through the utilization of counter balances and weight of the extremity with a cable and pulley arrangement. This technique gave load-assisting exercises to muscle groups that did not perform antigravity motions. McQueen (37) distinguished between exercise regimes for producing muscle hypertrophy and those for producing muscle power. He concluded that the number of repetitions for each set of exercise determines the different characteristics of the various training procedures. Figure 5 illustrates the complexity of the skeletal–muscular structure.

When muscles contract, the limbs may appear to move in unanticipated directions. One type of motion is a static contraction, known as an isometric type of contraction. Another type of contraction is a shortening or dynamic contraction that is called an isotonic contraction. Dynamic contractions are accompanied by muscle shortening and by limb movement. Dynamic contractions can exhibit two types of motion. One activity is a concentric contraction in which the joint angle between the two bones become smaller as the muscular tension is developed. The other action is an eccentric contraction in which, as the muscles contract, the joint angle between the bones increases. Owing to ambiguity in the literature concerning certain physiologic terms and differences in laboratory procedures, the following terms are defined below.

1. *Muscular strength*: the contractile power of muscles as a result of a single maximum effort.
2. *Muscular endurance*: ability of the muscles to perform work by holding a maximum contraction for a given length of time or by continuing to move submaximal load to a certain level of fatigue.

3. *Isometric*: a muscular contraction of total effort but with little or no visible limb movement (sometimes referred to as static or anaerobic).
4. *Isotonic*: a muscular contraction of less than total effort with visible limb movement (sometimes called dynamic or aerobic).
5. *Isokinetic training (accommodating resistance)*: muscular contraction at a constant velocity. In other words, as the muscle length changes, the resistance alters in a manner that is directly proportional to the force exerted by the muscle.
6. *Concentric contraction*: an isotonic contraction in which the muscle length decreases (that is, the muscle primarily responsible for movement becomes shorter).
7. *Eccentric contraction*: an isotonic contraction in which the muscle length of the primary mover and the angle between the two limbs increases during the movement.
8. *Muscle overload*: the workload for a muscle or muscle group that is greater than that to which the muscle is accustomed.
9. *Variable resistance exercise*: as the muscle contracts, the resistance changes in a predetermined manner (linear, exponentially, or as defined by the user).
10. *Variable velocity exercise*: as the muscle contracts with maximal or submaximal tension, the speed of movement changes in a predetermined manner (linear, exponentially, or as defined by the user).
11. *Repetitions*: the number of consecutive times a particular movement or exercise is performed.
12. *Repetition maximum (1 RM)*: the maximum resistance a muscle or muscle group can overcome in a maximal effort.
13. *Sets*: the number of groups of repetitions of a particular movement or exercise.

Based on evidence presented in these early studies (36–38), hundreds of investigations have been published relative to techniques for muscular development, including isotonic exercises, isometric exercises, eccentric contractions, and many others. The effectiveness of each exercise type has been supported and refuted by numerous investigations, but no definitive, irrefutable conclusions have been established.

Hellebrandt and Houtz (38) shed some light on the mechanism of muscle training in an experimental demonstration of the overload principle. They found that the repetition of contractions that place minimal stress on the neuromuscular system had little effect on the functional capacity of the skeletal muscles. They also found that the amount of work done per unit of time is the critical variable upon which extension of the limits of performance depends. The speed with which functional capacity increases suggests that the central nervous system, as well as the contractile tissue, is an important contributing component of training.

Results from the work of Hellebrandt and Houtz (38) suggest that an important consideration in both the design of equipment for resistive exercise and the performance of an athlete or a busy executive is that the human body relies on preprogrammed activity by the central nervous system. Since most human movements are ballistic and the neural control of these patterns differs from slow controlled movements, it is essential that training routines employ programmable motions to suit specific movements. This control necessitates exact precision in the timing and coordination of both the system of muscle contraction and the segmental sequence of muscular activity. Research has shown that a characteristic pattern of motion is present during any intentional movement of body segments against resistance. This pattern consists of reciprocally organized activity between the agonist and antagonist. These reciprocal activities occur in consistent temporal relationships with the variables of motion, such as velocity, acceleration, and forces.

In addition to the control by the nervous system, the human body is composed of linked segments, and rotation of these segments about their anatomic axes is caused by force. Both muscle and gravitational forces are important in producing these turning effects, which are fundamental in body movements in all sports and daily living. Pushing, pulling, lifting, kicking, running, walking, and all human activities result from the rotational motion of the links which, in humans, are the bones. Since force has been considered the most important component of athletic performance, many exercise equipment manufacturers have developed various types of devices employing isometrics and isokinetics. When considered as a separate entity, force is only one factor influencing successful athletic performance. Unfortunately, these isometric and isokinetic devices inhibit the natural movement patterns of acceleration and deceleration.

The three factors underlying all athletic performances and the majority of routine human motions are force, displacement, and the duration of movement. In all motor skills, muscular forces interact to move the body parts through the activity. The displacement of the body parts and their speed of motion are important in the coordination of the activity and are also directly related to the forces produced. However, it is only because of the control provided by the brain that the muscular forces follow any particular displacement pattern and, without these brain centre controls, there would be no skilled athletic performances. In every planned human motion, the intricate timing of the varying forces is a critical factor in successful performances. In any human movement, the accurate coordination of the body parts and their velocities is essential for maximizing performances. This means that the generated muscular forces must occur at the right time for optimum results. For this reason, the strongest weightlifter cannot put the shot as far as the experienced shotputter, although the weightlifter possesses greater muscular force, he has not trained his brain centers to produce the correct forces at the appropriate time. Older individuals may be unable to walk up and down stairs or perform many of the daily, routine functions that had been virtually automatic before the deterioration produced by weakness, disease, or merely age.

There are significant differences in the manner of execution of the various resistive training methods. In isotonic exercises, the inertia, which is the initial resistance, must be overcome before the execution of the movement progresses. The weight of the resistance cannot be heavier than the maximum strength of the weakest muscle acting in a particular movement or the movement cannot be completed. Consequently, the amount of force generated by the muscles during an isotonic contraction does not maintain maximum tension throughout the entire range of motion. In an isokinetically loaded muscle, the desired speed of movement occurs almost immediately and the muscle is able to generate a maximal force under a controlled and specifically selected speed of contraction.

The use of the isokinetic principle for overloading muscles to attain their maximal power output has direct applications in the fields of sport medicine and athletic training. Many rehabilitation programmed utilize isokinetic training to recondition injured limbs of athletes to their full range of motion. The unfortunate drawback to this type of training is that the speed is constant and there are no athletic activities that are performed at a constant velocity. The same disadvantage applies to normal human activities.

In isotonic resistive training, if more than one repetition is to be used, a submaximal load must be selected for the initial contractions in order to complete the required repetitions. Otherwise, the entire regimen would not be completed, owing to fatigue or, the inability to perform. A modality that can adjust the resistance so that it parallels fatigue to allow a maximum effort for each repetition would be a superior type of equipment. This function could be accomplished by manually removing weight from the bar while the subject trained. This is neither convenient nor practical. With the aid of the computer, the function can be performed automatically.

Another drawback with many isotonic types of resistive exercises is that the inertia resulting from the motion changes the resistance depending on the acceleration of the weight and of the body segments. In addition, since overload on the muscle changes due to both biomechanical levers and the length–tension curve, the muscle is able to achieve maximal overload only in a small portion of the range of motion. To overcome this shortcoming in resistive training, some strength training devices have been introduced that have "variable resistance" mechanisms, such as a cam, in them. However, these variable resistance systems increase the resistance in a linear fashion and this linearity may not truly accommodate the individual. When including inertial forces to the variable resistance mechanism, the accommodating resistance can be canceled by the velocity of the movement.

There seem to be unlimited training methods and each is supported and refuted by as many "experts". In the past, the problem of accurately evaluating the different modes of exercise was rendered impossible because of the lack of adequate diagnostic tools. For example, when trying to evaluate isotonic exercises, the investigator does not know exactly the muscular effort nor the speed of movement, but knows only the weight that has been lifted. When a static weight is lifted, the force of inertia provides a significant contribution to the load and cannot be quantified by feel or observation alone. In the isokinetic mode, the calibration of the velocity is assumed, but has been poorly verified since the mere rotation of a dial to a specific speed setting does not guarantee the accuracy of subsequently generated velocity. In fact, discrepancies as great as 40% have been observed when verifying the bar velocity.

Most exercise equipment currently available lack intelligence. In other words, the equipment is not aware that a subject is performing an exercise or how it is being conducted. Verification of the speed is impossible since a closed-loop feedback and sensors are absent. However, with the advent of miniaturized electronics in computers, it became possible to unite exercise equipment with the computer's artificial intelligence. In other words, it became possible for exercise equipment to adapt to the user rather than forcing the user to adapt to the equipment.

HIGH TECHNOLOGY TOOLS

High technology refers to the use of advanced, sophisticated, space age mathematical and electronic methods and devices for creating tools that can enhance human activities as well as expanding the horizons for future inventions. NASA put a man on the moon, sent exploratory spacecraft to Mars and beyond, and is sending shuttle missions to the Space Station. Polymer science invented plastics, mechanical science produced the automobile, and aeronautical engineering developed the airplane. Despite all of the knowledge and explosive developments since the rock became a tool, few advances have considered first the most important component in a complicated system, the human body.

The usual developmental cycle creates something and humans must adapt to it rather than the reverse. Computers can provide precise computations rapidly for complex problems that would otherwise require enormous quantities of time, talent, and energy to complete. The strength of these electronic wizards to follow instructions exactly, remember everything, and perform calculations within thousandths of a second has made them indispensable in finance, industry, and government. Application of the computer was a perfect enhancement for the human mind in order to quantify and evaluate movement performances. Used in conjunction with the human mind's ability to deduce, interpret, and judge, the computer provides the necessary enhancement to surpass the limits of what the eye can see or what intuition can surmise. Technological advances, such as these, can assist humans irrespective of their age.

For good health, it is necessary to follow a training method that incorporates all of the various bodily systems. In other words, the body should be treated as a complex, but whole, entity rather than as isolated parts. While it is not wrong to evaluate one's diet, an assessment of health would be incomplete without consideration of physical training, stress reduction, and other components that constitute the integrated organism of the human body. For a person to be able to jog 5 miles it is not important only to

run, but to develop the cardiovascular system in a systematic way to achieve a healthy status. Strength exercise, flexibility routines, proper nutrition and skill are necessary to achieve this goal.

Two sophisticated systems have been developed to analyze human performance and both are appropriate for the assault on aging. These systems include tools to (*1*) assess movements of the human body and (*2*) assist in exercising human beings. The first one is the biomechanical system that was developed to analyze movement performance. Currently, biomechanical analyses are routinely performed on a wide range of human motions in homes, work settings, recreation, hospitals, and rehabilitation centers. The second system, which incorporates space age technology, allows diagnoses and training of the musculoskeletal system. Each of these systems will be discussed subsequently in detail. Both of these technologies and the scientific principles and techniques discussed may help achieve physical and mental goals. The technological advances provide tools for quantification of the results and to analyze the potential of a person. With this information and these tools, it should be possible to train the various body systems for optimal results at any age.

The first commercially available computerized biomechanical system was described in 1973 (39) and that system can serve to illustrate the general concepts and procedures associated with biomechanical quantification of movement. Figure 6 illustrates device system. The computerized hardware–software system provides a means to objectively quantify the dynamic components of movement in humans, such as athletic events, gait analyses, work actions, as well as motion by inanimate objects, including such items as machinery actions, air bag activation, and auto crash dummies. This objective technique replaces mere observation and supposition. This system provides a means to quantity motion utilizing input information from any or all of the following mediums: visual (video), electromyography (EMG), force platforms, or other signal processing diagnostic equipment.

Figure 6. Analyses of vertical jump.

The Ariel Performance Analysis System provides a means of measuring human motion based on a proprietary technique for the processing of multiple high speed video recordings of a subject's performance (40–42). This technique demonstrates significant advantages over other common approaches to the measurement of human performance. First, except in those specific applications requiring EMG or kinetic (force platform) data, it is noninvasive. No wires, sensors, or markers need be attached to the subject. Second, it is portable and does not require modification of the performing environment. Cameras can be taken to the location of the activity and positioned in any convenient manner so as not to interfere with the subject. Activities in the workplace, home, hospital, therapist's office, health club, or athletic field can be studied with equal ease. Third, the scale and accuracy of measurement can be set to whatever levels are required for the activity being performed. Camera placement, lens selection, shutter and film speed may be varied within wide limits to collect data on motion of only a few centimeters or of many meters, with a duration from a few milliseconds to a number of seconds. Video equipment technology currently available is sufficiently adequate for most applications requiring accurate motion analysis. Determination of the problem, error level, degree of quantification, and price affect the input device selection.

A typical kinematic analysis consists of four distinct phases: data collection (filming); digitizing; computation; and presentation of the results. Data collection is the only phase that is not computerized. In this phase, video recordings of an activity are made using two or more cameras with only a few restrictions: (*1*) all cameras must record the action simultaneously. (*2*) If a fixed camera is used, it must not move between the recording of the activity and the recording of the calibration points. These limiting factors are not necessary when a panning camera and associated mechanism are used. A specialized device accompanied by specialized software was developed to accommodate camera movement particularly for use with gait analysis and some longer distance sporting events, such as skiing or long jumping. (*3*) The activity must be clearly seen throughout its duration from at least two camera views. (*4*) The location of at least six fixed noncoplanar points visible from each camera view (calibration points) must be known. These points need not be present during the activity as long as they can be seen before or after the activity. Usually they are provided by some object or apparatus of known dimensions that is placed in the general area of the activity, filmed and then removed. (*5*) The speed of each of the cameras (frames/second) must be accurately known, although the speeds do not have to be identical. (*6*) Some

Figure 7. Digitizing system.

event or time signal must be recorded simultaneously by all cameras during the activity in order to provide synchronization.

These rules for data collection allow great flexibility in the recording of an activity. Figure 7 illustrates a modern digitizing system to quantify human movement. Information about the camera location and orientation, the distance from camera to subject, and the focal length of the lens is not needed. The image space is self-calibrating through the use of calibration points that do not need to be present during the actual performance of the activity. Different types of cameras and different film speeds can be used and the cameras do not need to be mechanically or electronically synchronized. The best results are obtained when camera viewing axes are orthogonal (90° apart), but variations of 20–30° can be accommodated with negligible error. Initially, the video image is captured by the computer and stored in memory. This phase constitutes the "Grabbing" mode. Brightness, contrast, saturation, and color can be adjusted so that the grabbed picture may, in fact, be better than the original. Grabbing the image and storing it on computer memory eliminates any further need for the video apparatus.

Digitizing is the third step in biomechanical quantification. The image sequence is retrieved from computer memory and displayed, one frame at a time, on the digitizing monitor. Using a video cursor, the location of each of the subject's body joints (e.g., ankle, knee, hip, shoulder, elbow) is selected and stored in computer memory. In addition, a fixed point, which is a point in the field of view that does not move, is digitized for each frame as an absolute reference. The fixed point allows for the simple correction of any registration or vibration errors introduced during recording or playback. At some point during the digitizing of each view, a synchronizing event must be identified and, additionally, the location of the calibration points as seen from that camera must be digitized. This sequence of events is repeated for each camera view. This type of digitizing is primarily a manual process.

An alternative digitizing option permits the procedure to proceed automatically using any number of marker sets. This requires that the subject have the markers placed on the body prior to the filming phase. The types of markers and their placements have a substantial number of adherents particularly in the rehabilitation, gait measurement, and computer game communities. This type of digitizing combines manual and automatic, so that the activity progresses under manual control with computer-assisted selection of the joint segments or points. User participation in the digitizing process provides an opportunity for error checking and visual feedback which rarely slows the digitizing process adversely. A trained operator, with reasonable knowledge about digitizing and anatomy, can rapidly produce high quality digitized images. It is essential that the points are selected precisely because all subsequent information is based on the data provided in this phase.

The computation phase of analysis is performed after all camera views have been digitized. At this point in the procedures, the three-dimensional (3D) coordinates of the joints centers of a body are calculated. The transformation methods for transforming the data to two-dimensional (2D) or 3D coordinates are Direct Linear Transformation, Multiplier, and Physical Parameters Transformation. This phase computes the true 3D image space coordinates of the subject's body joints from the 2D digitized coordinates obtained from each camera's view. The Direct Linear Transformation Computation is determined by first relating the known image space locations of the calibration points to the digitized coordinate locations of those points. The transformation is then applied to the digitized body joint locations to yield true image space locations. This process is performed under computer control with some timing information provided by the user. The information needed includes, for example, starting and ending points if all the data are not to be used, as well as a frame rate for any image sequence that differs from the frame rate of the cameras used to record the sequence. The Multiplier technique for transformation is less rigorous mathematically and is utilized for those situations when no calibration device was used and only a few objects in the background are available to calibrate the area. This situation usually occurs when a nonscientific, third-party recorded the pictures such as a home video or even a televised sporting event. The third type of transformation, the Physical Parameters Transformation, is primarily applied with panning camera views or when greater accuracy is required on known image sources.

Following data transformation, a smoothing or filtering operation is performed on the image coordinates to remove small random digitizing errors and to compute body joint

velocities and accelerations. Smoothing options include polynomial, cubic and quintic splines, a Butterworth second-order digital and fast Fourier filters (43–45). Smoothing may be performed automatically by the computer or interactively with the user controlling the amount of smoothing applied to each joint. Error measurements from the digitizing phase may be used to optimize the amount of smoothing selected. Another unique feature is the ability to display the Power Spectrum for each of the x, y, and z coordinates. This enhancement permits the investigator to evaluate the effect of the smoothing technique and the chosen value selected for that curve by examining the Power Spectrum. Thus, the investigator can determine the method and level of smoothing that best meets the requirements of the specific research. After smoothing, the true 3D body joint displacements, velocities, and accelerations will have been computed on a continuous basis throughout the duration of the sequence.

Analogue data can be obtained from as many as 256 channels for input into the analogue-to-digital (A/D) system. Processing of the analogue signals, such as those obtained from transducers, thermistors, accelerometers, force platforms, EMG, ECG, EEG, or others, can be recorded for analysis and, if needed, synchronized with the video system. The displayed video picture and the vectors from the force plate can be synchronized so that the force vectors appear to be "inside the body". At this point, optional kinetic calculations can be performed to provide for measurement and analysis of the external forces that are applied to the body during movement. Inverse Dynamics are used to compute joint forces and torques as well as energy and momentum parameters of single or combined segments. External forces include anything external to the body that is applying force or resistance such as a golf club held in the hand. The calculations that are performed are made against the force distribution of the body.

The presentation phase of analysis allows computed results to be viewed and recorded in a number of different formats. Body position and motion can be presented in both still frame and animated stick figure format in 3D. Multiple stick figures may be displayed simultaneously for comparison purposes. Joint velocity and acceleration vectors may be added to the stick figures to show the magnitude and direction of body motion parameters. Copies of these displays can be printed for reporting and publication. Results can also be reported graphically. Plots of body joints and segments, linear and angular displacements, velocities, accelerations, forces, and moments can be produced in a number of format options. An interactive graphically oriented user interface allows the selection and plotting of such results to be simple and straightforward. In addition, body motion parameter results may also be reported in numerical form and printed as tables.

Utilizing this computerized system for biomechanical quantification of various movements performed by the elderly may assist in developing strategies of exercise, alterations in lifestyle, modifications in environmental conditions, and inventions to ease and/or extend independence. For example, rising from a chair is a challenging task for many elderly persons and getting up quickly is associated with a particularly high risk for falling. Hoy and Marcus (46) observed that older women moved more slowly and altered their posture to a greater extent than younger women. The strength levels were greater for the younger subjects, but it could not be concluded that strength was the causal mechanism for the slower speed. Following an exercise program affecting a number of muscle groups, younger and older women significantly increased in strength. Results of this study suggest that age-associated changes in muscle strength have an important effect on movement strategies used during chair rising. Following participation in a strength-training program, biomechanical assessment revealed changes in movement strategies that increased both static and dynamic stability. Other areas appropriate for biomechanical assessment would be on the well-known phenomenon of increased postural sway (47) and problems with balance (48–50) in the aged.

It is also important to study the motor patterns used by older persons while performing locomotor tasks associated with daily life such as walking on level ground and climbing or descending stairs. Craik (51) demonstrated that older subjects walking at the same speed as younger ones exhibited similar movement characteristics. Perhaps the older subjects selected slower movement speeds that produced apparent rather than real reductions in performance. These types of locomotors studies are easily assessed by biomechanical procedures. A biomechanical inquiry by Williams (52) examined the age-related differences of intralimb coordination by young and old individuals. Williams observed a similarity of general intralimb coordination for both old and young participants for level ground motions. One age-related change was suggested with regard to the additional balance constraints required for going up stairs because of adjustments not required on level ground. More profound differences were observed by Light et al. (53) with complex, multilimb coordinated movements performed in a standing position which necessitated dynamic balance control. These types of tasks showed significant age-dependent changes. Compared with younger subjects, the older participants were slower in all timing components, had less predominance in their movement patterns, less coupling of their limbs for movement end-points, and were more susceptible to environmental uncertainties. The alterations in movement performance reflected age-related loss in the ability to coordinate fast, multilimb movements performed from an upright stance suggesting that older individuals may have uncoordinated and unpredictable movement patterns when required to move quickly. Additionally, it was suggested that the more uncertain the environment, the greater the disturbance on the movement, thus, increasing the risk of falling. These studies provide realistic examples of one role biomechanics can perform by not only specifically identifying the locus of change but also providing objective quantification.

Another interesting application of the biomechanical system involves a multidimensional study of Alzheimer's disease currently in progress at a leading medical school. The study's strength is similar to the blind men who must integrate all of the information each has gathered in order to accurately describe the elephant. Examination of the

brain's response to specific drugs and at varying dosages, magnetic resonance imaging (MRI), thermographic, endocrine, and hormonal changes, vascular chemistry, as well as other aspects are being evaluated for each patient and their specific motor performances are being quantified biomechanically with the Ariel Performance Analysis system. Preliminary evidence indicates that performance on a simple bean-bag tossing skill improves daily although there is no cognitive recognition of the task. The activity of tossing a bean bag into a target circle from a standing position employs postural adjustments as well as coordinated arm and hand directed skills. Skill acquisition, or motor learning, involves both muscular capability and neural control mechanisms. Both activities involve closed- and open-loop mechanisms. The goal-directed movements needed to perform the bean-bag toss require the anticipatory postural adjustments that are inherent in an open-loop control. Because these findings suggest that muscular control and skill acquisition remain viable, this enables investigators to narrow the direction of the research and continue the study while continuously honing the focus. With each scientific finding, the research can be directed toward identification of the underlying cause.

The preceding discussion has described a computerized biomechanical system that can be utilized for the quantification of activities and performance levels particularly where appropriate for gerontological issues. Following the identification and definition of an activity, a second and equally necessary component follows. This is the ability to evaluate, test, and/or train the musculoskeletal components of the body in a manner appropriate to the specifically identified task(s) and according to the capabilities of the age and health of the individual. The integration of both technological assessment tools should assist the individual and others involved in their daily life to identify and measure those portions of an exercise program that can enhance performance, fitness status, or exercise capabilities for each gender and at different ages. In other words, one of the principles should be remembered is the goal of optimizing performance at every age.

For centuries, many devices have been created specifically for strength development. These devices include treadmills, bicycle ergometers, rowing machines, skiing simulators, as well as many of the more traditional resistive exercises with dumbbells, bar bells, and commercially available weight equipment. Figure 8 illustrates one of these equipment. Each type of exercise has some advantages, but none are designed to cope with the difficulties inherent with the gravitational effects that affect the multilinked human body performing on various exercise equipment.

All systems that employ weights as the mechanism for resistance have major drawbacks in four or more areas, as follows: (*1*) biomechanical considerations; (*2*) inertia; (*3*) risk of injury; (*4*) unidirectional resistance.

The biomechanical parameters are extremely important for human performance and should be incorporated into exercise equipment. The biomechanical factors were discussed previously. Inertia is the resistance to changes in motion. In other words, a greater force is required to begin moving weights than is necessary to keep them moving.

Figure 8. The computerized exercise equipment.

Similarly, when the exercising person slows at the end of a movement, the weights tend to keep moving until slowed by gravity. This phenomenon reduces the force needed at the end of a motion sequence. Inertia becomes especially pronounced as acceleration and deceleration increase, effectively reducing the useful range of motion of weight-based exercise equipment.

The risk of injury is obvious in most weight-based exercise equipment. When weights are raised during the performance of an exercise, they must be lowered to their original resting position before the person using the equipment can release the equipment and stop exercising. If the person exercising loses their grip, or is unable to hold the weights owing to exhaustion or imbalance, the weights fall back to their resting position; serious injuries can, and have, occurred. Finally, while being raised or lowered, weights, whether on exercise equipment or free standing, offer resistance only in the direction opposite to that of gravity. This resistance can be redirected by pulleys and gears but still remains unidirectional.

In almost every exercise performed, the muscle or muscles being trained by resistance in one direction are balanced by a corresponding muscle or muscles that could be trained by resistance in the opposite direction. With weight-based systems, a different exercise, and often a different mechanism, is necessary to train these opposing muscles. Exercise mechanisms that employ springs, torsion bars, and the like are able to overcome the inertia problem of weight-based mechanisms and, partially, to compensate the unidirectional force restriction by both expanding and compressing the springs. However, the serious problem of safety remains. An additional problem is the fixed, nonlinear resistance that is characteristic of springs and is usually unacceptable to most exercise equipment users.

The third resistive mechanism commonly employed in existing exercise equipment is a hydraulic mechanism. Hydraulic devices are able to overcome the inertial problem of weights, the safety problem of both weights and springs, and, with the appropriate selection or configuration, the unidirectional problem. However, previous applications of the hydraulic principle have demonstrated a serious

deficiency that has limited their popularity in resistive training. This deficiency is that of a fixed or a preselected flow rate through the hydraulic system. With a fixed-flow rate, it is a well established fact that resistance is a function of the velocity of the piston and, in fact, varies quite rapidly with changes in velocity. It becomes difficult for a person exercising to select a given resistance for training due to the constraint of moving either slower or faster than desired in order to maintain the resistance. Additionally, at any given moment, the user is unsure of just what the performing force or velocity actually is.

In the field of rehabilitation (54) especially, isokinetic or constant velocity training equipment is a technology that has enjoyed wide acceptance. These mechanisms typically utilize active or passive hydraulics or electric motors and velocity-controlling circuitry. The user or practitioner selects a constant level of velocity for exercise and the mechanism maintains this velocity while measuring the force exerted by the subject. Although demonstrating significant advantages over weight-based systems, isokinetic systems possess a serious limitation. There are virtually no human activities that are performed at a constant velocity. Normal human movement consists of patterns of acceleration and deceleration. When a person learns to run, ride a bike, or write, an acceleration–deceleration sequence is established that may be repeated at different rates and with different levels of force, but always with the pattern unique to that activity. To train, rehabilitate, or diagnose at a constant velocity is to change the very nature of the activity being performed and to violate most biomechanical performance principles.

FEEDBACK CONTROL OF EXERCISE

A newer form of exercise equipment can determine the level of effort by the person, compare it to the desired effort, and then adjust accordingly. The primary advantage of this resistive mechanism is that the pattern of resistance or the pattern of motion is fully programmable. The concept of applying a pattern of resistance or motion to training and rehabilitation was virtually impossible until the invention of computerized feedback control. Prior to the introduction of computerized feedback control, fitness technology could provide only limited modes of resistance and motion. Bar bells or weights of any type provide an isotonic or constant resistance type of training only when moved at a constant velocity. Typically, users are instructed to move the weights slowly to avoid the problem of inertia resulting from the acceleration or deceleration of mass. Weights used with cams or linkages that alter the mechanical advantage can provide a form of variable resistance. However, the pattern is always fixed and the varying mechanical advantage causes a variation in velocity that increases inertial effects. Users must move the weights slowly to preserve the resistance pattern. Another deficiency with these types of equipment is that they do not approximate the body or limb movement pattern of a normal human activity.

An exercise machine controlled by a computer possesses several unique advantages over other resistive exercise mechanisms, both fixed and feedback controlled. The most significant of these advances is the introduction of software to the human/computer feedback loop. The computer and its associated collection of unique programs can regulate the resistance to vary with the measured variables of force and displacement as well as modify the resistance according to data obtained from the feedback loop while the exercise progresses. This modification can, therefore, reflect changes in the pattern of exercise over time. The unique programmed selection can effect such changes in order to achieve a sequential or patterned progression of resistance for optimal training effect. The advantage of this capability over previous systems is that the user can select the overall pattern of exercise and the machine assumes responsibility for changing the precise force level, the speed of movement, and the temporal sequence to achieve that pattern.

There are a wide range of treadmills, bikes, and exercise devices currently available that employ electrical control features. These include such options as fat burn, up hill training, or cardiac modes. These types of equipment change the speed or elevations with preprogrammed actions that are determined at the manufacturing center when the machines are made rather than by the person exercising. The exerciser can select the programs presented on the control panel, but the response by the machine to the user is not at all related to the performance but rather to the preset events stored in the memory. Therefore, the person may be running "uphill" on the treadmill as determined by the imbedded system, but not with responsive interaction between the equipment and the individual moment by moment. This is a limitation of most of the exercise equipment available in the marketplace of the twenty-first century.

In the early 1980s, the first resistive training and rehabilitation device to employ computerized feedback control of both resistance and motion during exercise was introduced to overcome the lack of machine–human interactivity (55). For the first time, a machine dynamically adapted to the activity being performed rather than the traditional approach of modifying the activity to conform to the limitations of the machine. Biomechanical results previously calculated could be used to program the actual patterns of motion for training or rehabilitation. The equipment utilizes a passive hydraulic resistance mechanism operating in a feedback-controlled mode under control of the system's computer.

A simplified functional description of this mechanism, the Ariel Computerized Exercise System, and its operation is described. A hydraulic cylinder is attached to an exercise bar through a mechanical linkage. As the bar is moved, the piston in the hydraulic cylinder moves which pushes oil from one side of the cylinder, through a valve, and into the other side of the cylinder. When the valve is fully open there is no resistance to the movement of oil and, thus, no resistance in the movement of the bar. As the valve is closed, it becomes harder to push the oil from one side of the cylinder to the other and, thus, harder to move the bar. When the valve is fully closed, oil cannot flow and the bar will not move. In addition to the cylinder, the resistance mechanism contains sensors to measure the applied force on the bar and the motion of the bar. To describe

the operation of the computerized feedback loop, assume the valve is at some intermediate position and the bar is being moved at some velocity with some level of resistance. If the computer senses that the bar velocity is too high or that bar resistance is too low, it will close the valve by a small amount and then check the velocity and resistance values again. If the values are incorrect, it will continue to regulate the opening of the valve and continually check the results until the desired velocity or resistance is achieved. Similar computer assessments and valve adjustments are made for every exercise. Thus, an interactive feedback loop between the computer and the valve enable the user to exercise at the desired velocity or resistance. The feedback cycle occurs hundreds of times a second so that the user experiences no perceptible variations from the desired parameters of exercise.

There are a number of advantages in a computerized feedback controlled resistance mechanism over devices that employ weights, springs, motors, or pumps. One significant advantage is safety. The passive hydraulic mechanism provides resistance only when the user pushes or pulls against it. The user may stop exercising at any time and the exercise bar will remain motionless. Another advantage is that of bidirectional exercise. The hydraulic mechanism can provide resistance with the bar moving in each direction, whereas weights and springs provide resistance in only one direction. Opposing muscle groups can be trained in a single exercise. Two additional problems associated with weight training, noise and inertia, are also eliminated because the hydraulic mechanism is virtually silent and full resistance can be maintained at all speeds. Figure 9 illustrates an olympic training system utilized by the olympic athletes.

The Ariel Computerized Exercise System allows the user to set a pattern of continuously varying velocity or resistance. The pattern can be based on direct measurements of that individual's motion derived from the biomechanical analysis or can be designed or created by the user with a goal of training or rehabilitation. During exercise, the computer uses the pattern to adjust bar velocity or bar resistance as the subject moves through the full range of motion. In this manner, the motion parameters of almost any activity can be closely duplicated by the exercise system allowing training or rehabilitation using the same pattern as the activity itself.

The software consists of two levels. One level of software is invisible to the individual using the equipment since it controls the hardware components. The second level of software allows interaction between the user and the computer. The computer programs necessary to provide the real-time feedback control, the data program and storage, and the additional performance manipulations are extensive. The software provides computer interaction with the individual operator by automatically presenting a menu of options when the system is activated. Selection of the diagnostics option allows several parameters about that person to be evaluated and stored if desired. Some of the diagnostic parameters available include range of motion, maximum force, and maximum speed that the individual can move the bar for the specific activity selected. The maximum force and maximum speed data

Figure 9. Olympic training on the computerized exercise system.

can be determined at each discrete point in the range of movement as well as the average across the entire range. The diagnostic data can be used solely as isolated pre- and post-test measurements. However, the data can also be stored within the person's profile so that subsequent actions and tests performed on the equipment can be customized to adjust to that specific individual's characteristics.

The controlled velocity option permits the individual to control the speed of bar movement. The pattern of the velocity can be determined by the person using the equipment and these choices of velocity patterns include: (*1*) isokinetic, which provides a constant speed throughout the range of motion; (*2*) variable speed, in which the speed at the beginning of the motion and the speed at the end of the stroke are different with the computer regulating a smooth transition between the two values; and (*3*) programmed speed, which allows the user to specify a unique velocity pattern throughout the range of movement. For each of the choices, determination of the initial and final velocities is at the discretion of the individual through an interactive menu. The number of repetitions to be performed can be indicated by the person. Also, it is possible to designate different patterns of velocity for each direction of bar movement.

The controlled resistance option enables the person to control the resistance or amount of force required to move the bar. The alternatives include (*1*) isotonic, which provides a constant amount of force for the individual to overcome in order to move the bar; (*2*) variable resistance, in which the force at the beginning of the motion and the force at the end of the movement are different with the computer regulating a smooth transition between the two values; (*3*) programmed resistance, which permits the individual to specify a unique force pattern through-out the range of movement. An interactive menu enables the person to indicate the precise initial and final values, the number of repetitions to be used, and each direction of bar motion for the three choices. The controlled work option allows the individual to determine the amount of work, in Newton/meters or joules, to be performed rather than the number of repetitions. In addition, the person can choose either velocity or resistance as the method for controlling the bar movement. As with the previous options, bidirectional control is possible. The data storage capability is useful in the design of research protocols. The software allows an investigator to program a specific series of exercises and the precise manner in which they are to be performed, for example, number of repetitions and amount of work, so that the user need only select their name from the graphic menu and the computer will then guide the procedures. Data gathered can be stored for subsequent analysis. The equipment is fully operational for all options irrespective of whether the data storage option is activated.

Numerous features further enhance the application of this advanced fitness technology. Individual exercise programs can be created and saved on the computer, a CD, an internet file, or a USB disk. Users can perform their individual program at any time merely by loading it from any of the memory options used. Measurements of exercise results can be automatically saved and progress monitored by comparing current performance levels to previous ones. Performance can be measured in terms of strength, speed, power, repetitions, quantity of work, endurance and fatigue. Comparison of these quantities can be made for flexors versus extensors, right limb versus left limb, as well as between different dates and different individuals. Visual and audio feedback are provided during exercise to ensure that the subject is training in the proper manner and to provide motivation for optimal performance. Accuracy of measurement is essential and it is deemed as one of the most important considerations in the software. Calibration of the equipment is performed dynamically and is a unique feature that the computerization and the feedback system allow. Calibration is performed using weights with known values and the procedure can be performed for both up and down directions. This type of calibration is unique since the accuracy of the device can be ascertained throughout the range of motion.

FUTURE DEVELOPMENTS

As discussed previously, a large diagnostic and/or exercise system exists, but sheer bulk precludes its convenient use at home or in small spaces. One future goal is to develop a computerized, feedback-controlled, portable, battery-powered, hydraulic musculoskeletal exercise assessment and training equipment based on the currently available full-sized system. The device will be portable, compact, and operate at low voltage. Although physical fitness and good health have become increasingly more important to the American public, no compact, affordable, accurate device either for measurement or conditioning human strength or performance exists. This deficit hinders both America's ability to provide convenient, affordable, and accurate diagnostic and exercise capabilities for hospital or home-bound patients, children or elderly, to adequately perform within small-spaced military areas, as would be found in submarines, or in NASA shuttle projects to explore the frontiers of space. Figure 10 illustrates an astronaut running on a computerized treadmill in a zero gravity environment.

Figure 10. Motion analysis in space.

The frame will be compact and light-weight with a target weight of < 10 kg. This is an ambitious design goal that will require frame materials to have maximum strength/weight ratios and the structure must be engineered with attention directed toward compactness, storage size, and both ease and versatility of operation. The design of a smaller and lighter hydraulic valve, pack, and cylinder assembly is envisioned. Software can be tailored to specific applications such as for the very young or the aged, specific orthopedic and/or disease training, or other applications.

Another future development will be the ability to download programs through the Internet. For example, each patient could have one of the small exercise devices at home. His/her doctor can prescribe certain diagnostic activities and exercise regiments and transmit them via the Internet. The individual can perform the exercises at home and then submit the results to the doctor electronically. Biomechanical quantification of performances will become available electronically by downloading the software and executing the procedures on the individual's personal computer. Parents will be able to assist their child's athletic and

growth performances, doctors or physical therapists can compare normal gait with their patient's, and many other uses which may not be apparent at this time. The Internet can also function as a conduit between a research site and a remote location. Consider a hypothetical example of the National Institute of Health conducting a study on the effects of exercise on various medical, chemical, neural, and biomechanical factors for a large number of subjects around the world. The exercise equipment could be linked directly with Internet sources; the other data could be collected, and sent to the appropriate participating institutes. Findings from each location could then be transmitted to the main data collection site for integration.

CONCLUSION

National and international attitudes and policies focused on improving the health of children, workers, and the elderly must be directed towards good nutrition and improving lifestyles. It is made abundantly clean in print and televised media, that obesity has become a severe threat to the health and well being of Americans. That this problem is or will become an international epidemic may depend on the manner in which it is addressed. Exercise is no substitute for poor lifestyle practices, such as excessive alcohol consumption, smoking, overeating, and poor dietary practices. Attention must be directed to the importance of creative movements, posture, perceptual motor stimulation, body awareness, body image, and coordination. However, the importance of physical activity is too valuable to be limited to the young and healthy. Exercise, sports, and other physical activities must include all ages without regard to their frailty or disabilities.

The laws of nature rule the human body. Chemical and biological laws affect food metabolism, neurological transmissions within the nervous system and the target organs, hormonal influences, and all other growth, maintenance, and performance activities. Mechanical influences occur at the joints according to the same laws that return the pole vaulter to earth. Food, water, air, and environmental factors interact with work and societal demands. Human life is an interplay of external and internal processes and energy and, according to the second law of thermodynamics, the system will move toward increased disorder over time (56).

In terms of the universe, the first law of thermodynamics states that the total energy of the universe is constant. The second law states that the total entropy of the universe is increasing. The measure of a system's disorder is referred to as entropy and Eddington said, Whenever you conceive of a new theory of unusually attractiveness, but it does not in some way conform to the second law, then that theory is most certainly wrong (57). Everyone inevitably grows older. Delaying the process of disorder by keeping the subsystems of the organism at a low level of entropy does not flaunt the second law, but rather exploits it.

Science and technology have afforded us the ability to quantify movement so that humans can use their bodies more efficiently. Normal movement of small children can be reflected in improved diapers that do not alter their gait.

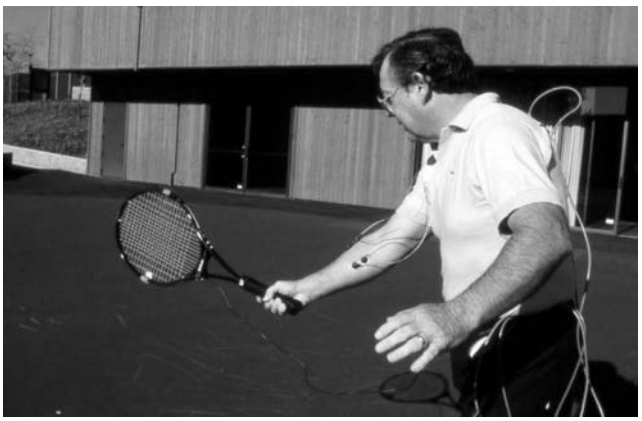


Figure 11. The EMG analysis of a tennis stroke.

Assessment of workplace activities can identify movements that are biomechanically inappropriate for healthy workers. Changing the design of the work bench, providing variable height stools for the conveyor belt operators, and evaluating the job requirements to assist in matching the employee to the work, improved wheelchair design, and adaptations in housing for the elderly are just a few examples of how biomechanical analysis can be applied. Figure 11 shows how athletic performance and equipment are assessed scientifically.

Not only has scientific and technological means provided quantitative assessment abilities, but has also allowed the development of improved means for exercising. Exercise equipment has become so sophisticated that it is appropriate for all ages. The youngest and the oldest can benefit from improved muscular health; the weakest and the strongest can always improve or, at the very least, sustain, healthy muscles; and those with compromised health or bodily functions should enjoy the opportunities to improve their musculature.

Logically, consumption of proper food, sleeping or resting sufficiently, and engaging in an appropriately amount of intense physical activity should keep the tissues and organs functioning maximally. To extend and improve the length and the quality of life depends on an increased awareness of human anatomy, biology, and physiology with continuous research efforts in these and other areas which impact human life. The aging process cannot be overcome, but it should be possible to negate many of the debilitating aspects of it. The Declaration of the United States of America is the only document of any country in history which includes the statement of "pursuit of happiness" and this concept should apply to the health and quality of life for all peoples, regardless of location, and at every age: from infancy to the twilight years.

BIBLIOGRAPHY

Cited References

1. Pollack L, Lowenthal DT, Graves JE, Carroll JF. The elderly and endurance training. In: Shephard RJ, Astrand P-O, editors. Endurance in Sport. London: Blackwell Scientific Publications; 1992. p 390–406.

2. Sidney K, Shephard R, Harrison J. Endurance training and body composition in the elderly. Am J Clin Nutr 1977;30:326–333.
3. Astrand P-O. Influences of biological age and selection. In: Shephard RJ, Astrand P-O, editors. Endurance in Sport. London: Blackwell Scientific Publications; 1992. p 285–289.
4. Bortz WM IV, Bortz WM II. Aging and the disuse syndrome - effect of lifetime exercise. In: Harris S, Harris R, Harris WS, editors. Optimization of human performance. Physical Activity, Aging and Sports, Vol. H.- P, Program and Policy. Albany (NY): Center for the Study of Aging; p 44–50.
5. Kraus H, Raab W. Hypokinetic diseases—diseases produced by the lack of exercise. Philadelphia: Thomas; 1961.
6. Bove AA. Heart and circulatory function in exercise. In: Lowenthal DT, Bharadwaja, Oaks WW, editors. Therapeutics Through Exercise. New York: Grune & Stratton; 1979. p 21–31.
7. Saltin B. et al. Response to exercise after bed rest and after training. Circulation 38(7): 1.
8. Erick H, Knottinggen A, Sarajas SH. Effects of physical training on circulation at and during exercise. Am J Cardiol 1963;12:142.
9. Saltin B, et al. Responses to exercise after bed rest and after training. Circulation 38(8): 1.
10. Clausen JP. Effect of physical training on cardiovascular adjustments to exercise. Ph Rev 1977;37:779.
11. Scheuer J, Tipton CM. Cardiovascular adaptations to physical training. Ann Rei Physiol 1977;39:221.
12. Ritzer TF, Bove AA, Lynch PR. Left ventricular size and performance followir term endurance exercise in dogs. Fed Proc 1977;36:447.
13. Miller PB, Johnson RL, Lamb LE. Effects of moderate exercise during four w(bed rest on circulatory function in man. Aerosp Med 1965;38:1077.
14. Oscai LB, Williams BT, Hertig BA. Effect of exercise on blood volume. J A Physiol 1968;24:622.
15. Hanson JS, Tabakin BS, Levy AM, Nedde W. Long term physical training and cardiovascular dynamics in middle aged men. Circulation 1968;38:783.
16. Becklake B, et al. Age changes in myocardial function and exercise response. Prog Cardiovasc Dis 1965;19:1–21.
17. Templeton G, Platt M, Willerson J, Weisfeldt M. Influence of aging on left ventricular hemodynamics and stiffness in beagles. Circ Res 1979;44:189–194.
18. Gottlieb SO, et al. Silent ischemia on Holter monitoring predicts mortality in high risk infarction patients. JAMA 1988;259:1030–1035.
19. Dustan H. Atherosclerosis complicating chronic hypertension. Circulation 1974;50:871.
20. Gribbin B, Pickering T, Sleight P, Peto R. Effect of age and high blood presst baroflex sensitivity in man. Circ Res 1971;29:424.
21. Bristow J, et al. Diminished baroflex sen; in high blood pressure. Circulation 1969;39:48.
22. Papper S. The effects of age in reducing renal function. Geriatrics 1973;28:83–87.
23. Lane C, et al. Long distance running, bone density, and osteoarthritis. JAMA 1986;255:1147–1151.
24. Shock N. Systems integration. In: Finch C, Hayflick L, editors. Handbook of the Bio1 Aging. New York: Van Nostrand Reinhold; 1977. p 639–665.
25. Cummings S, et al. Epidemiology of osteop and osteoporotic fractures. Epidemiol Rev 1985;7:178–208.
26. Basmajian JV, De Luca CJ. Muscles Alive. Baltimore: Williams & Wilkins; 1985.
27. Milliman and Robertson, Inc. Health risks and behavior: The impact on medical costs. C Data Corporation. 1987.
28. Lane N, et al. Long distance rut bone density and osteoarthritis. JAMA 1986;255:1147–1151.
29. Felig P, Koivisto V. The metabolic response to exercise: Implications for diabetes. In: Lowenthal DT, Bharadwaja K, Oaks WW, editors. Therapeutics Through Exercise. New York: Grune & Stratton; 1979. p 3–20.
30. Pollock M, Wilmore J, editors. Exercise in Health and Disease: Evaluation and Prescription for Prevention and Rehabilitation. Philadelphia: Saunders; 1990.
31. Jokl E. Physical activity and aging. In: Harris S, Harris R, Harris WS, editors. Physical Activity, Aging and Sports. Vol. II, Albany: Center for the Study of Aging; 1992. p 12–20.
32. Paffenbarger R, Hyde R, Wing A, Hsieh C. Physical activity and all-cause mortality and longevity of college alumni. N Engl J Med 1986;314:605–613.
33. Kannel W, et al. Prevention of cardiovascular disease in the elderly. J Am Coll Cardiol 1987;10:25A–8A.
34. Dudley G, Fleck S. Strength and endurance training: Are they mutually exclusive? Sports Med 1987;4:79–85.
35. McDonaugh M, Davies C. Adaptive response of mammalian skeletal muscle to exercise with high loads. Eur J Appl Phys 1984;52:139–155.
36. Delorme TL, Watkins AL. Techniques of progressive resistance exercise. Arch Phys Med 1948;29:645–667.
37. McQueen I. Recent advances in the technique of progressive resistance exercise. Br Med 1954;2:328–338.
38. Hellebrandt F, Houtz S. Mechanism of muscle training in man: Experimental demonstration of overload principle. Physiol Ther Rev 1956;36:371–376.
39. Ariel GB. Computerized biomechanical analysis of human performance. Mechanics and Sport, Vol. 4, New York: The American Society of Mechanical Engineers; 1973. p 267–275.
40. Wainwright RW, Squires RR, Mustich RA. Clinical significance of ground reaction forces in rehabilitation and sports medicine. Presented at the Canadian Society for Biomechanics, 5th Biannial Conference on Biomechanics and Symposium on Human Locomotion; 1988.
41. Llacera I, Squires RR. An analysis of the shoulder musculature during the forehand racquetball serve. Las Vegas: Presented at the American Physical Therapy Association meeting; 1988.
42. Susanka P. Biomechanical analyses of men's handball. Presented at International Handball World Federation 12th Men's Handball World Championship, Prague, Czechoslovakia, Charles University; 1990.
43. Reinsch C. Smoothing by spline functions. Numer Math 1967;10:177–183.
44. Wood GA, Jennings LS. On the use of spline functions for data smoothing. J Biomech 1975;12(6): 477–479.
45. Kaiser JF. Digital Filters. In: Liu D, editor. Digital Filters and the Fast Fourier Transform, 5-79. Stroudsburg (PA): Dowden, Hutchinson & Ross; 1975.
46. Hoy MG, Marcus R. Effects of age and muscle strength on coordination of rising from a chair. In: Woollacott M, Horak F, editors. Posture and Gait: Control Mechanisms. Vol. II. Portland (OR): University of Oregon Books; 1992. p 187–190.
47. Teasdale N, Stelmach GE, Bard C, Fleury M. Posture and elderly persons: Deficit; the central integrative mechanisms. In: Woollacott M, Horak F, editors. Posture and Gait: Control Mechanisms. Vol. II. Portland, (OR): University of Oregon Books; 1992. p 203–207.
48. Vamos L, Riach CL. Postural stability limits and vision in the older adult. In: Woollac M, Horak F, editors. Posture and Gait: Control Mechanisms. Vol. II. Portland (OR): University of Oregon Books; 1992. p 212–215.
49. Frank J, et al. Control of upright stand active, healthy elderly. In: Woollacott M, Horak F, editors. Posture and Gait:

Control Mechanisms. Vol. II. Portland (OR): University of Oregon Books; 1992. p 216–219.
50. Panzer V, Kaye J, Edner A, Holme L. Standing postural control in the elderly and v elderly. In: Woollacott M, Horak F, editors. Posture and Gait: Control Mechanisms. Vol. II. Portland (OR): University of Oregon Books; 1992. p 220–223.
51. Craik R. Changes in locomotion in the aging adult. In: Woollacott MH, Shumway-Co A, editors. Development of Posture and Gait across the Life Span. Columbia (SC): University of South Carolina Press; 1989. p 150–153.
52. Williams K. Intralimb coordination of older adults during locomotion: Stair climbing. In: Woollacott M, Horak F, editors. Posture and Gait: Control Mechanisms. Vol. II. Portland (OR): University of Oregon Books; 1992. p 208–211.
53. Light KE, Tang PF, Krugh CR. Performance differences between young and elderly females in a step-reach task. In: Woollacott M, Horak F, editors. Posture and Gait: Control Mechanisms. Vol. II. Portland (OR): University of Oregon Books; 1992. p 287–290.
54. Jacobs I, Bell DG, Pope J. Comparison of isokinetic and isoinertial lifting tests as predictors of maximal lifting capacity. Eur J Appl Physiol 1988;57:146–153.
55. Ariel GB. Computerized dynamic resistive exercise. In: Landry F, Orban WAR, editors. Mechanics of Sports and Kinanthropometry. Book 6. Miami (FL): Symposia Specialists, Inc.; 1978. p 45–51.
56. Benson H. University Physics. New York: John Wiley & Sons; 1991.
57. Eddington A. The Nature of the Physical World. Cambridge: New Press; 1928.

See also EXERCISE STRESS TESTING; HUMAN SPINE, BIOMECHANICS OF; JOINTS, BIOMECHANICS OF; LOCOMOTION MEASUREMENT, HUMAN; REHABILITATION AND MUSCLE TESTING.

BIOMECHANICS OF JOINTS. See JOINTS, BIOMECHANICS OF.

BIOMECHANICS OF SCOLIOSIS. See SCOLIOSIS, BIOMECHANICS OF.

BIOMECHANICS OF SKIN. See SKIN, BIOMECHANICS OF.

BIOMECHANICS OF THE HUMAN SPINE. See HUMAN SPINE, BIOMECHANICS OF.

BIOMECHANICS OF TOOTH AND JAW. See TOOTH AND JAW, BIOMECHANICS OF.

BIOMEDICAL ENGINEERING EDUCATION

PAUL BENKESER
Georgia Institute of Technology
Atlanta, Georgia

INTRODUCTION

Biomedical engineering is that interdisciplinary field of study combining engineering with life sciences and medicine. It is a relatively new field of study that has only recently experienced sufficient maturity to enable it to establish its own identity. Often, this field will be described using the term bioengineering. In 1997, the Bioengineering Definition Committee of the National Institutes of Health released the following definition of the field (1): "Bioengineering integrates physical, chemical, mathematical, and computational sciences and engineering principles to study biology, medicine, behavior, and health. It advances fundamental concepts; creates knowledge from the molecular to the organ systems level; and develops innovative biologics, materials, processes, implants, devices and informatics approaches for the prevention, diagnosis, and treatment of disease, for patient rehabilitation, and for improving health."

While many use biomedical engineering and bioengineering interchangeably, it is generally accepted today that bioengineering is a broader field that combines engineering with life sciences, but is not necessarily restricted to just medical applications.

The Biomedical Engineering Society further elaborated on the definition of biomedical engineering as part of a guide on careers in the field. In it is stated (2): "A Biomedical Engineer uses traditional engineering expertise to analyze and solve problems in biology and medicine, providing an overall enhancement of health care. Students choose the biomedical engineering field to be of service to people, to partake of the excitement of working with living systems, and to apply advanced technology to the complex problems of medical care. The biomedical engineer works with other health care professionals including physicians, nurses, therapists and technicians. Biomedical engineers may be called upon in a wide range of capacities: to design instruments, devices, and software, to bring together knowledge from many technical sources to develop new procedures, or to conduct research needed to solve clinical problems."

Educational programs in the field of biomedical engineering had their origins in a handful of specialized graduate training programs in the 1950s focusing primarily on diagnostic and therapeutic devices and instrumentation. By 2004, there were undergraduate and graduate programs in biomedical engineering at ~100 universities in the United States. The diversity in the content of undergraduate educational programs that was commonplace in its early years is gradually diminishing as the field has matured. While the current undergraduate programs still maintain their own unique identity, there has been a steady movement toward the definition of a core curriculum in the field.

The purpose of this article is to give the reader some historical perspective on the origins of educational programs in the field, the challenges associated with preparing bachelor-level graduates for careers in the field, and the current state-of-the-art in undergraduate biomedical engineering curriculums.

HISTORY

The first steps toward establishing biomedical engineering as a discipline occurred in the 1950s as several formalized training programs were created. Their establishment was significantly aided by the National Institutes of Health

creation of training grants for doctoral studies in biomedical engineering. The Johns Hopkins University, the University of Pennsylvania, the University of Rochester, and Drexel University were among the first to be awarded these grants.

During the late 1960s and early 1970s, growing opportunities in the field helped prompt the development of a second generation of biomedical engineering programs and departments. These included Boston University in 1966; Case Western Reserve University in 1968; Northwestern University in 1969; Carnegie Mellon University, Duke University, Rensselaer Polytechnic Institute and a joint program between Harvard and the Massachusetts Institute of Technology in 1970; Ohio State University and University of Texas, Austin, in 1971; Louisiana Tech, Texas A&M and the Milwaukee School of Engineering in 1972; and the University of Illinois, Chicago in 1973 (3). Many of these first and second generation of programs were concentrating the training of their students in areas defined either using quasiclassical engineering terminology, such as bioinstrumentation, biomaterials and biomechanics, or by application area, such as rehabilitation engineering or clinical engineering.

The late 1990s witnessed a substantial increase in the growth of the number of departments and programs in biomedical engineering, especially at the undergraduate level. The growth of this third generation of programs was fueled in part by grants from The Whitaker Foundation to help institutions establish or develop biomedical engineering departments or programs. In 2004, ~100 universities have programs or departments in biomedical engineering, including 33 offering undergraduate degree programs accredited by the Engineering Accreditation Commission of the Accreditation Board for Engineering and Technology (ABET) (4). The growth in the numbers of ABET accredited degree programs is illustrated in Fig. 1.

The arrival of the third generations of programs coincided with the development of several new areas of training in biomedical engineering, such as systems biology–physiology, and tissue, cellular, and biomolecular engineering. These areas typically require significantly more training in life sciences than was present in the first and second generation biomedical engineering training programs. This presented significant challenges for undergraduate programs trying to add this life science content to their curricula without increasing the number of credit hours required for the programs. Many of these programs accomplished this by creating a new generation of courses in which the engineering and life science concepts are integrated together within courses. The integration of such courses into the curriculum is discussed in more detail in the Curriculum section.

CAREER PREPARATION

The design of a high quality educational program should always start with its educational objectives. By using the definition established by ABET, these program educational objectives are statements that describe the expected accomplishments of graduates during the first several years following graduation (5). This requires programs to be cognizant of the needs of prospective employers of its graduates and design learning environments and curricula to meet those needs. This is particularly challenging task for a relatively new and evolving field like biomedical engineering.

Biomedical engineers are employed in industry, in research facilities of educational and medical institutions, in teaching, in government regulatory agencies, and in hospitals. They often serve as integrators or facilitators, using their skills in both the engineering and life science fields. They may work in teams in industry to help design devices, systems, and processes that require an in-depth understanding of both living systems and engineering. Frequently, biomedical engineers will be found in technical sales and marketing positions in companies seeking to provide their customers with technically trained individuals who are capable of better understanding their needs and communicating those needs back to product development teams. Government regulatory positions, such as those with the Food and Drug Administration, often involve testing medical devices for performance and safety. In research institutions, biomedical engineers participate in or direct research activities in collaboration with other

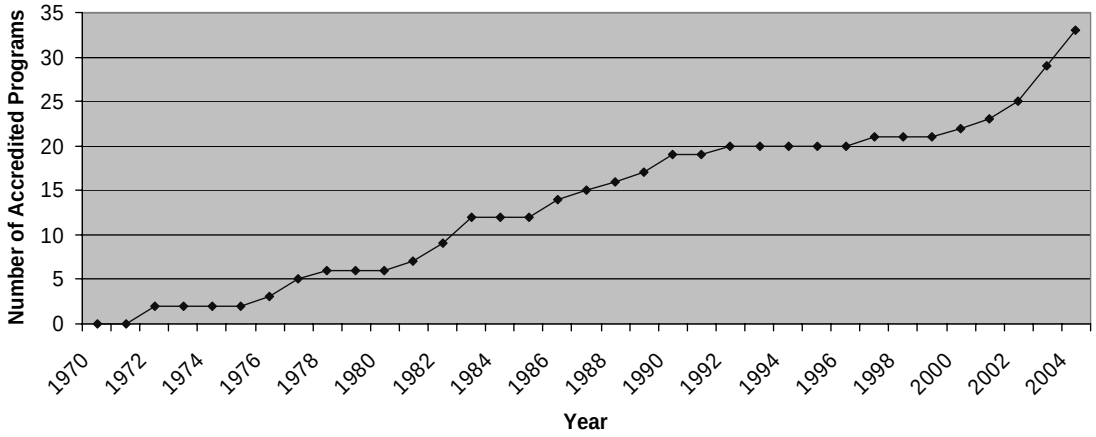


Figure 1. Number of ABET accredited programs in biomedical enginering.

researchers with such backgrounds as medicine, biology, chemistry, and a variety of engineering disciplines.

According to the U.S. Department of Labor's Bureau of Labor Statistics, manufacturing industries employed 38% of all biomedical engineers, primarily in the pharmaceutical and medicine manufacturing and medical instruments and supplies industries (6). Others worked in academia, hospitals, government agencies, or as independent consultants. Employment of biomedical engineers is expected to increase faster than the average for all occupations through 2012 (6). The demand for better and more cost-effective medical devices and equipment designed by biomedical engineers is expected to increase with the aging of the population and the associated increased focus on health issues. Most of these employment opportunities will be filled with graduates from B.S. and M.S. degree programs. However, for research-oriented jobs, like faculty positions in academia and research and development positions in industry, employers typically require their employees to possess a Ph.D. degree.

The needs of the employers that hire biomedical engineers undoubtedly vary by industry and job title. However, there are some skills that appear to be in universal demand by all employers of biomedical engineers. They include proficient oral and written communication skills, the ability to speak the languages of engineering and medicine, a familiarity with physiology and pathophysiology, and teamwork skills (6).

In spite of this seemingly impressive list of career paths and options, one of the most significant challenges facing entry-level biomedical engineers are prospective employers who complain that they do not understand what skill sets biomedical engineers possess (7). The perception is that the skills possessed by engineers from other disciplines, like electrical or mechanical engineering, are more predictable and in large part are independent of the university where the engineer was educated. It is likely that this perception is a result of a combination of two factors. First, often the individuals responsible for making the hiring decisions in companies did not receive their degrees in biomedical engineering, and thus do not have first-hand experience with the training received by biomedical engineers. Second, until relatively recently, many undergraduate biomedical engineering programs lacked a substantive core curriculum and were structured in such a way that students had to select from one of several "tracks" offered by the program. These tracks were typically pattered along traditional engineering lines, such as bioelectronics and biomechanics, in an attempt to address another concern expressed by prospective employers—that graduates of bachelor degree programs in biomedical engineering were too broadly trained and thus lacked sufficient depth of engineering skills. Due to this perceived lack of depth, it is not uncommon to find employers for which the entry-level degree for biomedical engineering positions is the masters degree. Undoubtedly the presence of these tracks, and their variability from program to program, contributed to the confusion in industry over the what skill sets they should expect from a biomedical engineer.

As a result of these concerns over depth, breadth, and uniformity of curriculum, the biomedical engineering education community has recognized the need to reach consensus on what constitutes a core undergraduate curriculum in biomedical engineering. This has become one of the major initiatives of the National Science Foundation (NSF) sponsored VaNTH Engineering Research Center (ERC) for Biomedical Engineering Educational Technologies. This ERC, a collaboration of teams from Vanderbilt University, Northwestern University, The University of Texas at Austin, Harvard University, and the Massachusetts Institute of Technology, was created in 1999. Its vision is to transform bioengineering education to produce adaptive experts by developing, implementing and assessing educational processes, materials, and technologies that are readily accessible and widely disseminated (8). The Whitaker Foundation has also sponsored workshops at its 2000 and 2005 Biomedical Engineering Summit meetings with the goal of delineating the core topics in biomedical engineering that all biomedical engineering students should understand. White papers from these meetings can be found at the Foundation's web site (9).

In spite of the movement toward the creation of a common core curriculum in undergraduate programs of study in biomedical engineering, there will undoubtedly continue to be some differences in curricula between programs. This is not only permitted in the current accreditation review process of ABET, but in some sense encouraged. Within the past decade this process has changed from a prescriptive evaluation to an outcomes-based assessment centered on program-defined missions and objectives (10). Thus, it will be incumbent on programs to work closely with the prospective employers of their graduates to ensure that the programs provide the graduates with the skills the employers desire. For example, Marquette University's biomedical engineering program has an established industrial partners program with >30 companies participating (11).

THE UNDERGRADUATE CURRICULUM

Contained within this section is a description of a core undergraduate curriculum that the author believes the biomedical engineering educational community is converging upon. The contents are based upon reviews of curriculums from biomedical engineering programs (12) and information disseminated by the Curriculum Project of the VaNTH ERC (13). It is important to note that the core described herein is not being presented as the prescription for what a biomedical engineering curriculum should look like, but rather a reflection of the current trends in the field.

Course Requirements

To be accredited by ABET, the curriculum must include the following:

- One year of an appropriate combination of mathematics and basic sciences.
- One-half year of humanities and social sciences.
- One and one-half years of engineering topics and the requirements listed in ABETs Program Criteria for bioengineering.

A year is defined as 32 semester or 48 quarter hours.

The typical math and science content of biomedical engineering curriculums, as described in Table 1, are similar to those of other engineering disciplines. The notable exceptions are courses in biology and organic chemistry. In general, most programs rely on other departments within their university to provide the instruction for these courses. Some programs with large enrollments have been successful in collaborating with faculty in their university's science departments to create biology and chemistry courses for biomedical engineering students. For example, the School of Chemistry and Biochemistry at the Georgia Institute of Technology has created organic and biochemistry courses specifically designed for biomedical engineering students.

Table 1. Mathematics and Science Core Curriculum Subjects

Subject	Sampling of Topical Coverage
Math	Linear algebra
	Differential, integral, multivariable calculus
	Differential equations
	Statistics
Physics	Classical mechanics, oscillations and waves
	Electromagnetism, light and modern physics
Computer Science	Algorithms, data structures, program design and flow control
	Graphics and data visualization
	Higher level programming language (e.g., Matlab, Java, C++)
Chemistry	General and inorganic chemistry
	Organic chemistry
	Biochemistry
Biology	Modern biological principles
	Genetics
	Cell biology

The core biomedical engineering content is described in Table 2. Depending on the size of the program, some of the content may be delivered in courses outside of biomedical engineering (e.g., thermodynamics from mechanical engineering). It is not uncommon to find some variability between programs in the content of their core curriculums. This will likely always be the case as each program must provide the curriculum that best enables its graduates to achieve the program's unique educational objectives.

ABET Criterion 3 stipulates that engineering programs must demonstrate achievement of a minimum set of program outcomes. These outcomes are statements that describe skills that "students are expected to know or be able to do by the time of graduation from the program" (5). A closer examination of these skills suggests that they can be divided into two sets as illustrated in Table 3. The first set, "domain" skills, is one that engineering educators are typically adept in both teaching and quantitatively measuring achievement. Programs generally use courses, like those listed in Tables 1 and 2, to develop these domain

Table 2. Biomedical Engineering Core Curriculum Subjects

Subject	Sampling of Topical Coverage
Biomechanics	Principles of statics
	Mechanics of biomaterials
	Dynamics
Biotransport	Mass transfer
	Heat transfer
	Momentum transfer
Biothermodynamics	Thermodynamic principles
	Mass and energy balances
Biomaterials	Metals, polymers and composite materials
	Biocompatibility
Bioinstrumentation	Instrumentation concepts
	Amplifiers and filters
	Sensors and transducers
Biofluids	Blood vessel mechanics
	Hydrostatics and steady flow models
	Unsteady Flow and non-uniform geometric models
Systems Physiology	Cellular metabolism
	Membrane dynamics
	Homeostatis
	Endocrine, cardiovascular and nervous systems
	Muscles
Biosignal Analysis	Digital signal processing theory
	Filtering
	Frequency-domain characterization of signals

skills in their students. The second set, "professional" skills, is more difficult to teach and assess. However, these professional skills are often the ones most frequently cited by employers of engineers as the most important skills they value in their employees.

Humanities and social science courses are integral to the achievement of these professional skills. However, programs must avoid employing the "inoculation" model for teaching these skills to the students. In this model, it is assumed that students can learn these skills by simply taking isolated courses in ethics, technical communications, and so on. There are several problems with this model. It can decontextualize these skills, treating them as add-ons and not an integral part of everyday engineering practice. This is a false and even dangerous message to give the students—that written and oral communication and ethical behavior are peripheral to the real world of engineering. This message is further driven home because the faculty responsible for teaching these skills is humanities or social science faculty not engineering faculty. In addition, the complexity of these skills to be learned is too great for students to master within the framework of isolated courses. Research suggests, however, that students need quasirepetitive activity cycles and practice in multiple settings to develop proficiency in these professional skills (14–16).

Professional Skills

Before describing methods of developing these professional skills in students, it is necessary to establish operational

Table 3. Program Outcomes Specified in ABET Criterion 3

Domain Skills	Professional Skills
An ability to apply knowledge of mathematics, science, and engineering	An ability to function on multi-disciplinary teams
An ability to design and conduct experiments, as well as to analyze and interpret data	An understanding of professional and ethical responsibility
An ability to design a system, component, or process to meet desired needs	An ability to communicate effectively
An ability to identify, formulate, and solve engineering problems	The broad education necessary to understand the impact of engineering solutions in a global and societal context
A knowledge of contemporary issues	A recognition of the need for, and an ability to engage in lifelong learning
An ability to use the techniques, skills, and modern engineering tools necessary for engineering practice	

descriptions for these constructs. Such descriptions serve two functions. They reveal the complexity of the particular skills in terms of the subskills required to demonstrate the higher level skills specified in the ABET lists. Descriptions can also serve as articulations of learning outcomes, which can be designed toward and assessed. The following represents one interpretation of the variables that are indicators of these constructs (16).

Ability to communicate effectively.

Oral + written communication skills

- Convey information and ideas accurately and efficiently.
- Articulate relationships among ideas.
- Inform and persuade.
- Assemble and Organize evidence in support of an argument.
- Make communicative purpose clear.
- Provide sufficient background to anchor ideas–information.
- Be aware of and address multiple interlocutors.
- Clarify conclusions to be drawn from information.

Ability to function on multidisciplinary teams.

Team − collaboration skills + communication skills 3

- Help group develop and achieve team goals.
- Avoid contributing excessive or irrelevant information.
- Confront others directly when necessary.
- Demonstrate enthusiasm and involvement.
- Monitor group progress and complete tasks on time.
- Facilitate interaction with other members.

Understanding professional and ethical responsibilities.

- Recognize moral problems and issues in engineering.
- Comprehend, clarify, and critically assess opposing arguments.
- Form consistent and comprehensive viewpoints based on facts.
- Develop imaginative responses to problematic conflicts.
- Think clearly in the midst of uncertainty and ambiguity.
- Appreciate the role of rationale dialogue in resolving moral conflicts.
- Ability to maintain moral integrity in face of pressures to separate professional and personal convictions.

Broad education necessary to understand the impact of engineering solutions in a global and societal context.

- Identify human needs or goals technology will serve.
- Analyze and evaluate the impact of new technologies on economy, environment, physical and mental health of manufacturers, uses of power, equality, democracy, access to information and participation, civil liberties, privacy, crime and justice.
- Identify unintended consequences of technology development.
- Create safeguards to minimize problems.
- Apply lessons from earlier technologies and experiences of other countries.

Recognition of the need for, and an ability to engage in life-long learning.

- Identify learning needs and set specific learning objectives.
- Make a plan to address these objectives.
- Evaluate inquiry.
- Assess the reliability of sources.
- Evaluate how the sources contribute to knowledge.
- Question the adequacy and appropriateness of forms of evidence used to report back on learning needs.
- Apply knowledge discovered to the problem.

Table 4. Repetitive Activities in the Problem-Solving Cycle

	Professional Skill				
Activity	Communicate	Teams	Responsibilities	Impact	Learning
Identifying learning–knowledge needs as a team–individual					X
Acquiring knowledge needed to solve problem			X	X	X
Reporting back to team	X	X			X
Digging deeper and solving	X	X	X	X	X
Presenting solution to audience of experts	X				
Writing a report on problem solution	X				

There are a variety of methods programs can employ to foster the development of these professional skills in students. These include the use of team-based capstone design experiences, facilitating student participation in coop and internship experiences, encouraging involvement in undergraduate research projects, and incorporating oral and written communication exercises throughout the curriculum.

The creation of new undergraduate biomedical engineering programs has led to the development of some new approaches to professional skills development. For example, the Wallace H. Coulter Department of Biomedical Engineering at Georgia Tech and Emory University has implemented an integrative approach to the development of professional skills and adaptive expertise in its program. This approach anchors professional skills development in the context of team-based problem solving and design experiences over the four-year curriculum. This approach provides multiple opportunities for the students to work on and develop skills and knowledge in a variety of "real-world" engineering settings in which these professional skills are practiced. Within each experience, activities are identified within the problem-solving cycle that help students develop these professional skills. These activities are described in Table 4.

The need for real-world problems cannot be overstated. Authentic, open-ended problems are needed as contexts and catalysts for the development of these professional skills. Not only do they help prepare students for the professional practice of engineering, but they are also a significant motivator for the students to delve more deeply into the problem space. Moreover, the use of these skills in the context of a large problem makes them central not peripheral to biomedical engineering problem solving. They begin to understand the value of clear, thoughtful communication, and collaboration when confronting complex problems. They see how ethical issues can arise when seeking design solutions. If the problems are authentic, then the information needed to solve them must be found in multiple places, which helps students to develop inquiry and research skills for lifelong learning.

SUMMARY

The field of biomedical engineering had its foundations laid roughly 50 years ago. Undergraduate degree programs in the field followed shortly thereafter. Fueled in part by generous support from the Whitaker Foundation, there has been a significant increase in the number of new undergraduate degree programs in the field. This has led to a significant increase in student interest in the field. This growth has increased the need for the biomedical engineering education community to work with industry to better define the skills that graduates need to obtain to lead productive careers in the field. There exists a movement, led by the NSF VaNTH ERC, to define a core undergraduate curriculum within the constraints imposed by ABET accreditation criteria. The VaNTHs vision to transform bioengineering education to produce adaptive experts has been adopted by new undergraduate degree programs and has produced demonstrated examples of pedagogical advances in the field of engineering education.

BIBLIOGRAPHY

Cited References

1. NIH working definition of bioengineering. 1997, July 24. National Institutes of Health Bioengineering Consortium. Available at http://www.becon.nih.gov/bioengineering_definition.htm. Accessed 2004 Nov. 18.
2. Planning a career in biomedical engineering. 1999. Biomedical Engineering Society. Available at http://www.bmes.org/careers.asp. Accessed 2004 Nov. 18.
3. A history of biomedical engineering. 2002, May. The Whitaker Foundation. Available at http://www.whitaker.org/glance/definition.html. Accessed 2004 Nov. 19.
4. Accredited engineering programs. 2004. Accreditation Board for Engineering and Technology. Available at http://www.abet.org/ accredited_programs/engineering/EACWebsite.asp. Accessed 2004 Nov. 19.
5. Criteria for accrediting engineering programs. 2004. Accreditation Board for Engineering and Technology. Available at http://www.abet.org/criteria.html. Accessed 2004 Nov. 19.
6. Bureau of Labor Statistics, U.S. Department of Labor, Occupational Outlook Handbook. 2004–2005 edition, Biomedical Engineers. Available at http://www.bls.gov/oco/ocos262.htm. Accessed 2005 Feb. 10.
7. RA Linsenmeier, What makes a biomedical engineer, IEEE Eng Med Biol Mag 2003;22(4):32–38.
8. Cordray DS, Pion GM, Harris A, Norris P. The value of the VaNTH Engineering Research Center. IEEE Eng Med Biol Mag 2003;22(4):47–54.
9. Biomedical Engineering Educational Summit, The Whitaker Foundation. Available at http://summit.whitaker.org/. Accessed 2005 Feb 10.
10. Enderle J, Gassert J, Blanchard S, King P, Beasley D, Hale P, Aldridge D. The ABCs of preparing for ABET. IEEE Eng Med Biol Mag 2003;22(4):122–132.

11. Waples LM, Ropella KM. University partnerships in biomedical engineering. IEEE Eng Med Biol Mag 2003;22(4): 118–121.
12. The biomedical engineering curriculum database 2004. The Whitaker Foundation. Available at http://www.whitaker.org/academic/database/index.html. Accessed 2004 Nov. 18.
13. VaNTH ERC curriculum project (2004, June 10). VaNTH ERC [Online]. Available at http://www.vanth.org/curriculum/. Accessed [2004 Nov. 18].
14. Bransford JD, Brown AL, Cocking RR, editors. How People Learn: Brain, Mind, Experience, and School. Washington: National Academy Press; 1999.
15. Harris TR, Bransford JD, Brophy SP. Roles for learning sciences and learning technologies in biomedical engineering education: A review of recent advances. Annu Rev Biomed Eng 2002;4:29–48.
16. Benkeser PJ, Newstetter WC. Integrating soft skills in a BME curriculum, Proc. 2004 ASEE Annu Conf. 2004, June. American Society for Engineering Education. Available at http://www.asee.org/about/events/conferences/search.cfm. Accessed 2004 Nov. 19.

See also BIOINFORMATICS; MEDICAL EDUCATION, COMPUTERS IN; MEDICAL ENGINEERING SOCIETIES AND ORGANIZATIONS.

BIOSURFACE ENGINEERING

PETER MOLNAR
MELISSA HIRSCH-KUCHMA
JOHN W. RUMSEY
KERRY WILSON
JAMES J. HICKMAN
University of Central Florida
Orlando, Florida

INTRODUCTION

One primary reason there is a tremendous amount of interest in cellular patterning techniques is that numerous examples in nature use these techniques to segregate cells into tissues, vessels, and organs. The idea of templates in nature abounds for the creation of organized biological systems using both inorganic (1), as well as organic template systems (2). There is also a certain allure to being able to integrate electrically active cells directly to electronic devices using standard electronic fabrication techniques. Researchers have attempted to use surface cues to pattern cells since as early as 1917, in which spider webs were used to pattern cells (3). Most of the early work on cellular patterning used topographical cues until 1988, when a landmark publication by Kleinfeld et al. (4) used lithographic templating to fabricate simple patterns of cortical neurons. This was an adaptation of standard technology developed by the electronics industry to create computer chips that was then applied to the creation of patterns to guide neuronal cell attachment. It was at this point that interest in this field mushroomed, as the idea of creating neuronal networks from living neurons has potential applications in understanding biological information processing, creating hybrid computer systems, as well as a whole host of biomedical applications. Some prominent efforts to use cell patterning have been for spinal cord repair, creation of *in vitro* test bed systems to study diseases, as well as blood vessel formation from patterned endothelial cells (5). The initial lithography-based technique used by Kleinfeld et al. has been extended to include many other methods for the patterning of cells, including self-assembled monolayer (SAM) patterning, laser ablation, microcontact printing or stamping, ink jet printing, AFM printing, as well as patterning using microfluidic networks. Methods have also been extended in the area of topographical cues for 3D patterning, which has evolved from early work that used scratched grooves in glass surfaces. At this point, depending on the facilities that one has available, some form of mask making and pattern templating is available to just about any laboratory in the world. However, the biological interactions with these patterns that have been created are still not well developed, and this limits applications at this point.

There are many reasons for the lack of long-term applications of this technique, even after the large amount of work that has been done in this area. The first is that, in many instances, the patterns direct the initial attachment of cells, but as the extracellular matrix is deposited by the cells, long-term adherence to the patterns is not maintained. Another issue is that longer term cell survival is also dependent on factors besides the surface, such as media composition, cell–cell contact interactions, and the lack of growth factors that are normally present from other support cells and tissues. Defined systems are being developed in an attempt to control these other variables in addition to the surface (6,7), but these efforts have been limited to date. However, as progress is made in these areas, it will open up possible applications in tissue engineering, tissue repair, biosensors, and functional *in vitro* test beds.

PATTERNING METHODS

Many methods have been developed for creating templates to be used for cell patterning. These can be divided roughly into two categories: those that are derived from photolithography techniques and those that depend on physical segregation, although there is some crossover in the methods between the two. The photolithography-based systems typically use some sort of organic layer, from polymers to monolayers, which is illuminated, either directly with a pattern or through the template pattern to be created. This can involve many or a few steps depending on the particular method used. Generally, a second layer is deposited in the area where material was removed. Specific variations of this technique are discussed in this section.

The second major category, physical segregation, involves the actual placement of the molecules or cells in a pattern on a surface. Stamping is the most well-known method of creating a molecular-based template and of all the techniques is probably the most economical, but other techniques have been investigated for physical placement of cells in a desired pattern on the surface. Finally, both of these methods, which are 2D in nature, are now being extended into 3D patterning using many of the same

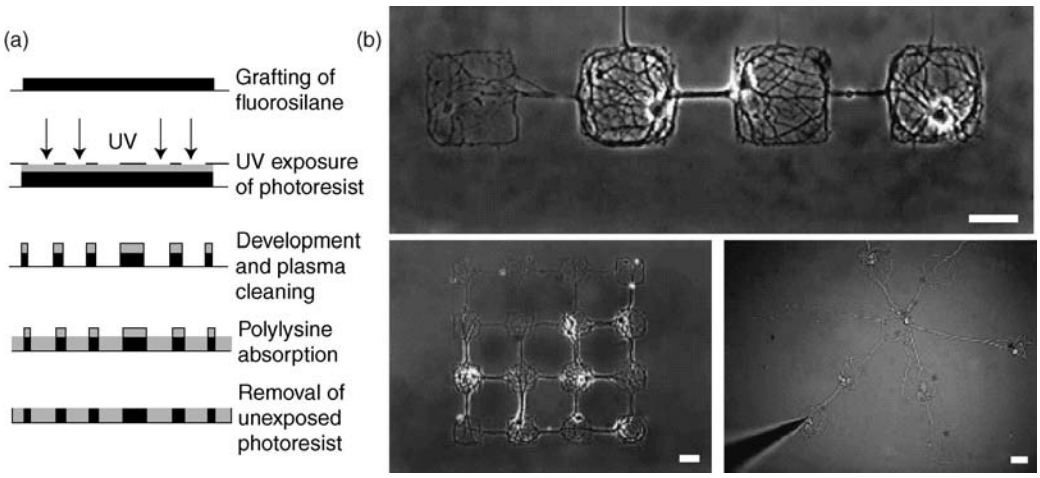


Figure 1. Protocol and images of the pattern. (a) Protocol of photolithography. A glass coverslip (continuous line) is coated with domains of fluorosilane molecule (dark) and with regions of polylysine (light gray) according to the pattern designed on the mask (dashed line), using ultraviolet (UV) exposure of a spin-coated photoresist (gray). (b) Images of neural networks of controlled architecture. Top: linear network. Bottom left: matrix 4×4. Bottom right: star. Cell bodies of neurons are restricted to squares or disks of 80 μm and neurites to line (80 μm length, 2–4 μm wide). Square and disk diameters are 80 μm for each figure. Scale bar is 50 μm (Ref. 9).

techniques, or combinations thereof, described therein. Below there is a brief description, along with the appropriate references, of the techniques that can be used to create these cellular templates.

Photolithography

A variety of photolithographic techniques has been employed to pattern proteins and cells on surfaces from the micrometer to the nanometer scale (8). The basic tools needed are a radiation source and a photomask. The photomask can be created using standard photolithography processes developed for the electronics industry. Irradiation of the surface through the photomask is used to create the patterns by ablation or by using a photosensitive material such as a photoresist as shown in Fig. 1.

Kleinfeld et al. (4) first demonstrated that dissociated neurons could be grown on 2D substrates consisting of lithographically defined patterned monolayers of diamines and triamines with alkylsilanes. The method used by Kleinfeld et al. started with a clean silicon or quartz surface that was spin coated with a layer of photoresist (an organic photosensitive polymer used in the electronics industry). The resist was exposed to UV light with a patterned photomask and then developed. The surface was refluxed in the presence of an alkylsilane, and the photoresist was then stripped off so that areas the photomask covered were reduced to bare silicon or quartz. These areas were then reacted with an aminosilane to form the patterned surface. The patterned cells developed electrical excitability and immunoreactivity for neuron-specific proteins. A further modification of this technique eliminated the photoresist from the pattern formation by direct ablation of the SAM layer (10).

Patterns of self-assembled monolayers formed from organosilanes on glass or silicon substrates and on gold surfaces can be made by using a photoresist mask and deep UV radiation (10–18). Monolayers can also be directly ablated with various forms of radiation such as UV, X ray, ions, and electrons (8) depending on the resolution needed for the patterns. Organosilanes self-assemble and condense onto substrates that have surface –OH functionalities (19). The –SH functionality of the alkanethiol (20) is also highly reactive to ozone and other irradiation sources and has been used in patterning (21). Methods using X-ray or extreme UV (EUV) radiation give better resolution than the traditional photolithography using deep UV and photoresist masks. The ablated regions of the SAM can then be reacted with an organosilane or alkanethiol with different characteristics from the original layer to enable cell growth (22). Azides (23) and aromatic hydrocarbon silanes (24) have also been shown to be reactive for creating patterns.

A typical method to prepare a patterned glass silanated surface is illustrated next. The glass must first be acid cleaned or oxygen plasma cleaned to maximize the surface –OH functional density. Next, the glass is reacted with a silane that contains -chloro, -methoxy, or -ethoxy bonds in the presence of a small amount of water that acts as a catalyst. A mask is then used to protect certain areas of the surface while allowing the radiation source to ablate others in the desired pattern. The ablated regions of the surface can then be coated with a different silane with different properties than the original, thus forming the patterned surface; an example of this is shown in Fig. 2.

The photoreactivity of polymers, such as poly(ethylene) glycol and polystyrene, has also been used to pattern surfaces (25). Biologically based polymers, such as poly-L-lysine and extracellular matrix proteins, have been ablated tso create patterns (26). Polymer photolithography followed by protein adsorption has been combined to create patterned cytophobic and cytophilic areas. Patterning of perfluoropolymers followed by adsorption of poly-L-lysine

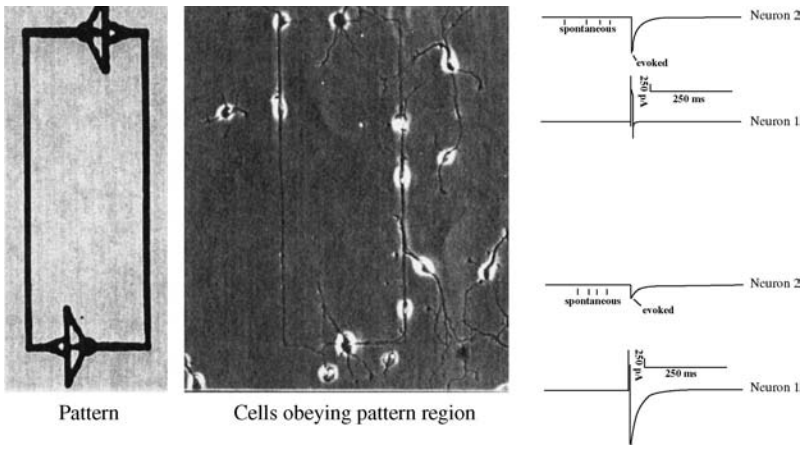


Figure 2. Micrograph of circuit-patterned day 2 *in vitro* hippocampal neurons plated onto DETA/15F modified glass coverslips. Electrophysiology of day 12 *in vitro* hippocampal neurons displaying both spontaneous and evoked activity on a DETA/15F line-space patterned surface. Top trace: post-synaptic neuron. Bottom trace: stimulated presynaptic neuron.

and albumin has been one method used (27). Photosensitive polymers have been treated with UV radiation to create an –COOH functional unit on the surface and then patterned via the linkage of proteins to form cytophilic and cytophobic regions (28). A bioactive photoresist (bioresist) has been developed that does not require the use of solvents that can denature the biomolecule that is patterned (29).

Photolithographic protein patterning requires the attachment of photosensitive groups to proteins on a substrate. Patterns can be made using a patterned mask and selective ablation. This method has been shown to be useful to produce micropatterned cultures (30). Various methods are used to covalently attach proteins in patterns to surfaces (31), to pattern using biomolecule photoimmobilization (32), and to create density gradients of photoreactive biomolecules (33). Heterobifunctional crosslinker molecules have been used to attach proteins to silanated surfaces both before and after the photolithographic patterning step (34). Protein patterning has been achieved using a micromirror array (MMA), which can transfer a pattern from the mirrors that are switched on, ablating a photolabile protecting group (35). Photolithography was also used to pattern thermosensitive copolymers through polymer grafting (36). The surface micropattern appeared and disappeared interchangeably, as observed under a phase-contrast microscope, by varying the temperature between 10 and 37 °C. The copolymer-grafted polystyrene surface was hydrophobic at 37 °C and hydrophilic at 10 °C.

Photolithography provides high resolution patterning and the ability to make complex patterns on surfaces. Unlike stamping techniques, the patterns are more permanent; however, the process can be relatively expensive as it generally requires the use of a laser and clean room facilities for the mask production. To attach proteins, such as ECM proteins, the use of a covalently attached crosslinker is necessary, and stamping techniques are generally preferred for this application.

Microcontact Printing (Stamping)

Microcontact printing was introduced by George Whiteside's group at Harvard in 1994 (37) to pattern self-assembled monolayers on gold substrates to control surface properties, cell adhesion, proliferation, and protein secretion by patterned cells. The basic method to create surface patterns by microcontact printing has not changed much since then. Usually, a poly(dimethylsiloxane) (PDMS) stamp is created using a molding technique from a master pattern relief mold and then used to transfer chemical patterns to flat or curved surfaces. The master is usually prepared from silicon by standard photolithography and/or etching, but other substrates can also be used. The transferred chemical patterns can be created using a compound that binds covalently to the substrate (e.g., self-assembled monolayers or proteins immobilized by crosslinkers) (38) or a compound that binds noncovalently, such as absorbed extracellular matrix proteins (39). This methodology is illustrated in Fig. 3. Oliva et al. presented a novel method to couple proteins to patterned surfaces based on the strong interaction of protein A and the Fc fragment of immunoglobulins. This method involved the creation of a covalently coupled Fc fragment and the target protein (41). Methods have also been developed to transfer proteins from a fluid phase to a surface using hydrogels as the stamp (42). Moreover, recently introduced techniques are allowing the creation of protein gradients with microcontact printing (43). Although alignment of the stamp/patterns with surface features such as microelectrodes is more difficult than in the case of photolithographic patterning, several groups are beginning to address this issue (44). Microcontact printing is usually a favored method among biologists compared with photolithography, because (*1*) the equipment and controlled environment facilities required for photolithography are not routinely available to cell biologists and (*2*) the steps are simpler to pattern proteins, the molecules of greatest interest to biologists, using microcontact printing than with photolithography and crosslinkers. The refinement of the PDMS molding technique has directly led to the development of another important patterning method, microfluidics, which gained wider applications with the introduction of microelectromechanical systems (MEMS) and "lab-on-a-chip" systems. However, initial results using PDMS indicated some transfer of the PDMS to the surface during the stamping process. This can be troublesome in cell patterning applications as PDMS can be toxic to cells or mask the chemical functionality of interest. Methods of "curing" the stamps or presoaking to enable better release of the compound has been reported (45).

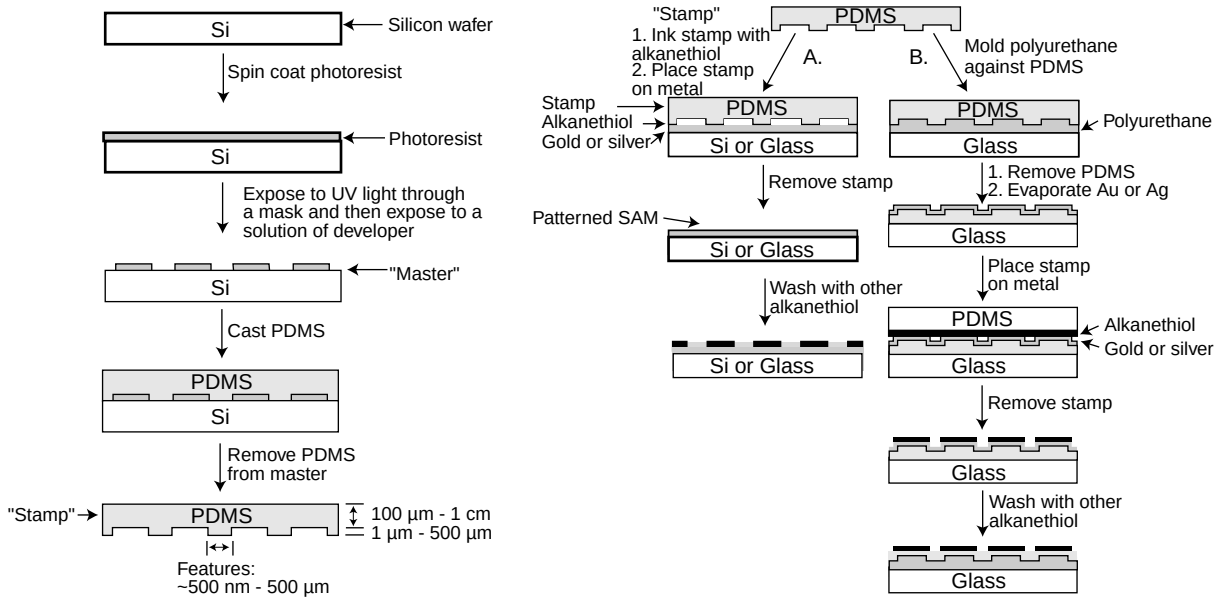


Figure 3. The creation of the master stamp is indicated on the left side of the figure, and its use to make patterns is indicated on the flow chart to the right (40).

Inkjet Printing

Inkjet technology used in desktop printers can be applied to the creation of viable cell patterns by printing proteins on to a surface (46). It is a fast and inexpensive method that does not require any contact (such as with stamping) to the surface. This method is desirable for high throughput printing of surfaces, and there is good control of the drop volume and of the alignment of the pattern. Printing occurs when small volumes of a protein solution or a solution containing cells is pumped through a nozzle in a continuous jet or small droplets of the solution are formed either by an acoustic or thermal pulse. The drop size is in the range of 10–20 pL (8). Inkjet printing and the use of computer-aided design (CAD) have impacted the biomaterials field greatly in the areas of biosensor development (47), immobilization of bacteria on biochips (48), DNA arrays and synthesis (49), microdeposition of proteins on cellulose (50), and free-form fabrication techniques to create acellular polymeric scaffolds. A drawback of this method is that the resolution, in the 20–50 μm range, is limited by the statistical variations in the drop direction and spreading on the surface (8). Other high throughput printing methods have also been adapted to pattern proteins on surfaces for biological application such as the already developed DNA spotter used to create DNA microchips (51).

Patterning via Microfluidic Networks

Synthetic surfaces may also be patterned using microfluidic networks (μFNs) to selectively generate regions with greater cytophilicity. This method involves the use of a microfluidic network fabricated in an elastomeric polymer, usually polydimethylsiloxane (PDMS), to direct a protein solution to the regions where cell-adhesion is desired. Gravity and pressure-driven flows are the most common methods for circulating the solution.

The most basic application of this method involves allowing a solution of the material to be patterned noncovalently to the substrate surface using the microchannels as guides. Some of the earliest work done with this method by Folch and Toner (52) involved patterning of various polymer surfaces with human plasma fibronectin (Fn) and collagen to create adhesion promoting domains for hepatocyte/fibroblast cocultures. Similar work by Chiu et al. (53) involved the patterning of glass coverslips with fibrinogen (Fb) and bovine serum albumin (BSA) for patterned cocultures with bovine adrenal capillary endothelial cells (BCEs) and human bladder cancer cells (ECVs). Further work by Takayama et al. (54) demonstrated an added degree of sophistication of this technique by using the laminar flow characteristics of microchannels to generate patterns using a single microchannel.

In addition to simple noncovalent binding of proteins to the substrate, it is possible to use a crosslinking agent to covalently link a molecule of interest. Delamarche et al. (55) used a hydroxylsuccinimidyl ester to chemically couple immunoglobulin G (IgG) to various substrates. Another more commonly used method involves functionalized silane SAMs on substrates, such as 3-aminopropyltriethoxysilane (APTES), and a crosslinking reagent, such glutaraldehyde (GA), to achieve crosslinking of protein molecules to a substrate. Romanova et al. (25) demonstrated the applicability of this method using microfluidic patterning to study controlled growth of *Aplysia* neurons on geometric patterns of poly-L-lysing and collagen IV. Yet another variation on this method was developed by Itoga et al. (56), who generated patterns by photopolymerization of acrylamide on 3-methacryloxypropyltrimethoxysilane modified glass coverslips. In this method, the acrylamide monomer was flowed through microchannels adhered to the derivatized coverslip and cross-linked via photopolymerization to generate cytopholic poly-acrylamide regions.

Perhaps the most sophisticated application of microfluidic patterning of substrates for cell adhesion was demonstrated by Tan and Desai (57,58), who demonstrated the fabrication of complex multilayer cocultures for biomimetic blood vessels. In this work, the 3D structure of blood vessels was recreated by differential deposition of protein and cell layers on a glass coverslip. By alternately layering proteins (Collagen I, collagen/chitosan, and matrigel) and cell types found in blood vessels (fibroblasts, smooth muscle, and endothelial cells), it was possible to recreate layers mimicking the adventitial, medial, and intimal layers observed in blood vessels.

Topographical (3D) Patterning Methods

Control of Cell Placement, Movement, and Process Growth Based on Topographical Clues. It has been known for many years that cells react to topographical clues in their environment (59). Originally, natural fibers were used to create topological clues. Later, fabrication methods developed for the microchip industry were adapted to micromachine silicon surfaces for cell culture applications (45). For the creation of micro- or nanotopography, sophisticated methods and equipment are necessary, which are available in most electronics laboratories, but are usually not available to cell biologists. In recent years, as a response to the increased need for high- throughput screening methods (planar patch clamp, lab-on-a-chip) and in response to the challenge of biological applications of nanoscience, several multidisciplinary team/centers have been established with microfabrication capabilities. The most commonly used fabrication methods are (*1*) silicon etching (60), (*2*) photoresist-based methods (61), (*3*) PDMS molding (62), and (*4*) polymer/hydrogel molding (63). Methods have also been developed for the creation of complex 3D structures by the rapid prototyping/layer-by-layer technique (58). Tan et al. (62) used 3D PDMS structures not only to control attachment and morphology of cells but also to measure attachment force through flexible microneedles as the culture substrate. Xi et al. used AFM cantilevers to demonstrate and measure the contracting force of cardiac muscle cells (64).

3D Patterning of Living Cells. Several hydrogel scaffold-based methods have also been developed to create 3D patterns from cells. For example, photo-polymerization of hydrogels can be used to create patterns of entrapped cells (65). These scaffolding methods offer possible intervention in spinal cord injury (66). Layer-by-layer methods have also been used to create complex "tissue analog" cellular structures, such as blood vessels (57).

Other Patterning Methods

Several other methods, based on microcontact printing and PDMS stamping, have been developed to create cellular patterns. Folch et al. used an inexpensive method to create microwells for coculture experiments based on a reusable elastomeric stencil (i.e., a membrane containing thru holes), which seals spontaneously against the surface (67,68). Gole and Sastry developed a novel method to pattern surfaces with lipids followed by selective protein incorporation into the lipid patterns that result in complex protein patterns on the surface (69). A scanning electrochemical microscope has also been adapted to pattern self-assembled monolayers on surfaces with high spatial resolution by either chemical removal of SAMs (70) or by gold deposition (71). Amro et al. used an AFM tip to directly print nanoscale patterns using the so-called dip-pen technique (72–74). However, none of these methods has been proven beyond the demonstration step.

Applications

Many *potential* applications for cell patterning remain in the biomedical and biotechnology fields. However, there has been limited success to date, besides demonstrations in the literature, for any of the applications envisioned by the host of researchers in this area. However, the promise of the use of cellular patterning for real applications is bolstered by the success that has been achieved for patterning of DNA, RNA, and protein arrays (42,51) as well as for enzymatic biosensors, such as simple pregnancy tests. Much like with cellular patterning, there was a period of time during the development of molecular patterning techniques before the applications became relevant, and the authors believe this is the situation that exists for cell patterning at this time. One reason for this long development stage is that viable and reproducible cellular patterns have many other variables that are not a major issue with biomolecule patterning applications. The cellular media are very important for cell survival, especially long term, as well as cell preparation, which has a significant affect on an extracellular attachment. In addition, no universal combination will be good for every cell, as each cell type has a unique environment that it needs to survive and function, and some aspect of these factors needs to be reproduced for long-term applications. That said, many examples of tissues exhibit some segregation, including blood vessels, lung tissue, the lining of the stomach and intestines, as well as a host of other tissues, which could benefit from this methodology. However, one of the most studied cell tissues is that of the central nervous system (CNS), which exhibits a complicated network of structures that will be difficult to reproduce for reconstruction or repair of neuronal tissue or for other *in vivo* as well as *in vitro* applications. To date, there has been some success in manipulating cells in patterns and controlling certain variables that would be necessary for the creation of functional tissues, but a complete system, using this methodology, has not been reported in the literature. However, there has been some success in demonstrating the intermediate steps that will be necessary for the realization of applications of this method for biomedical and biotechnological applications. Cell attachment has been demonstrated by several researchers, and generally, pattern adherence is maintained for approximately 1 week although longer times have been demonstrated (75). Control of cell morphology and differentiation are two important factors that are necessary in creating functional systems, and these have also been demonstrated. For neuronal-based systems, the primary variable, that of axonal polarity, has also been demonstrated. Brief descriptions and progress in these areas are described below.

Controlling Cell Attachment, Morphology, and Differentiation. *In vivo*, cells are arranged in distinct patterns (76). This patterning effect is dictated during development, with cues provided by both physical contact with other cells and chemicals present in the extracellular matrix (77). Because the random arrangement of cells cultured *in vitro* does not represent the complex architecture seen in tissues, studies of many cell types lack a clear *in vivo* relationship. Consequently, techniques to create defined and reproducible functional patterns of cells on surfaces have been created.

Controlling Attachment and Morphology. Several methods have been developed to control the attachment of cells on surfaces creating patterns that more accurately mimic conditions found *in vivo*. Using photolithography, microcontact printing, and microstamping, groups have been able create 2D patterns that guide cell attachment and alignment (37,78–81). Three-dimensional patterning techniques have also been employed to influence cell orientation and polarity of neurons and osteoblasts (82–84). Cells have also been attached to capillaries and microfluidic devices using SAMs and protein adsorption or microcontact printing (85,86). Additionally, cell attachment and proliferation has been enhanced using biomolecules attached covalently, by stamping, or microcontact printing to surfaces (37,78–82,84–87).

Controlling Morphology and Differentiation of Cell Types Other Than Neurons. Microfabrication and photolithography used to create microtextured membranes for cardiac myocyte culture showed greater levels of attachment and cell height relative to 2D culture techniques (88). Similar techniques applied to vascular smooth muscle also showed the ability to control shape and size of the cells (89). Furthermore, microtextured surfaces were shown to influence gene expression and protein localization in neonatal cardiomyocytes (90). Cues provided to cells by the topography of their extracellular environment are thought to play a role in differentiation. The generation of microtopographical surfaces in titanium has been used to regulate the differentiation of osteoblasts *in vitro* (91).

Study of Axon Guidance in Neurons. Using photolithographic techniques and SAMs, the 2D patterns created were shown to influence neuronal polarity (22). Photolithographically fabricated 3D surfaces demonstrated that topology also influenced the orientation of neurons and the polarity of axonal outgrowth (83). The ability to guide neurite outgrowth and axonal elongation has significant applications in the areas of spinal cord repair, synapse, formation, and neural network formation. Initial studies using striped patterns on glass coverslips showed that neurons would adhere and preferentially extend axons along the length of the pattern (4,92). Growth of neurons on micropatterned 2D surfaces showed preferential axon extension along the length of the pattern as well as increased axon extension (92–96).

The use of 3D microchannels and microstructured surfaces has also been shown to increase the complexity of neuronal architecture, increase neurite growth, and enhance cell activity (61). Photolithography has also been used to pattern neurons and control axon elongation for the formation of neuronal networks (25,97). The synapses formed by these hippocampal neurons showed strong electrophysiological activity up to 17 days in culture (10). These network formations show promise for use in screening pharmacological agents as well as for electronic connection.

BIBLIOGRAPHY

Cited References

1. Fritz M, Belcher AM, Radmacher M, Walters DA, Hansma PK, Stucky GD, Morse DE, Mann S. Flat pearls from biofabrication of organized composites on inorganic substrates. Nature Biotechnol 1994;371:49–51.
2. Noctor SC, Flint AC, Weissman TA, Dammerman RS, - Kriegstein AR. Neurons derived from radial glial cells establish radial units in neocortex. Nature 2001: 409(6821): 714–720.
3. Harrison RG. The reaction of embryonic cells to solid structures. J Exp Zool 1914;17:521–544.
4. Kleinfeld D, Kahler KH, Hockberger PE. Controlled outgrowth of dissociated neurons on patterned substrates. J Neurosci 1988;8(11):4098–4120.
5. Spargo BJ, Testoff MA, Nielsen TB, Stenger DA, Hickman JJ, Rudolph AS. Spatially controlled adhesion, spreading, and differentiation of endothelial-cells on self-assembled molecular monolayers. Proc Nat Acad Sci 1994;91(23):11070–11074.
6. Das M, Molnar P, Gregory C, Riedel L, Jamshidi A, Hickman JJ. Long-term culture of embryonic rat cardiomyocytes on an organosilane surface in a serum-free medium. Biomaterials 2004;25(25):5643–5647.
7. Das M, Bhargava N, Gregory C, Riedel L, Molnar P, Hickman JJ. Adult rat spinal cord culture on an organosilane surface in a novel serum-free medium. In Vitro Animal Cell Develop Bio 2005. In press.
8. Geissler M, Xia Y. Patterning: Principles and some new developments. Adv Mater 2004;16(15):1249–1269.
9. Wyart C, Ybert C, Bourdieu L, Herr C, Prinz C, Chatenay D. Constrained synaptic connectivity in functional mammalian neuronal networks grown on patterned surfaces. J Neurosci Methods 2002;117(2):123–131.
10. Dulcey CS, Georger JHJr. Krauthamer V, Stenger DA, Fare TL, Calvert JM. Deep UV photochemistry of chemisorbed monolayers: Patterned coplanar molecular assemblies. Science 1991;252(5005):551–554.
11. Dressick WJ, Calvert JM. Patterning of self-assembled films using lithographic exposure tools. Appl Phys Part 1 1993; 32(12B):5829–5839.
12. Bhatia SK, Teixeira JL, Anderson M, Shriver-Lake LC, Calvert JM, Georger JH, Hickman JJ, Dulcey CS, Schoen PE, Ligler FS. Fabrication of surfaces resistant to protein adsorption and application to two-dimensional protein patterning. Anal Biochem 1993;208(1):197–205.
13. Liu J, Hlady V. Chemical pattern on silica surface prepared by UV irradiation of 3-mercaptopropyltriethoxy silane layer: Surface characterization and fibrinogen adsorption. Colloids Surfaces B-Biointerfaces 1996;8(1–2):25–37.
14. Dressick WJ, Dulcey CS, Chen MS, Calvert JM. Photochemical studies of (aminoethylaminomethyl)phenethyltrimethoxysilane self-assembled monolayer films. Thin Solid Films 1996;285:568–572.
15. Georger JH, Stenger DA, Rudolph AS, Hickman JJ, Dulcey CS, Fare TL. Coplanar patterns of self-assembled monolayers for selective cell-adhesion and outgrowth. Thin Solid Films 1992;210(1–2):716–719.

16. Stenger DA, Georger JH, Dulcey CS, Hickman JJ, Rudolph AS, Nielsen TB, McCort SM, Calvert JM. Coplanar molecular assemblies of aminoalkylsilane and perfluorinated alkylsilane—characterization and geometric definition of mammalian-cell adhesion and growth. J Am Chem Soc 1992;114 (22):8435–8442.
17. Calvert JM. Lithographic patterning of self-assembled films. J Vacuum Sci Technol B 1993;11(6):2155–2163.
18. Ravenscroft MS, Bateman KE, Shaffer KM, Schessler HM, Jung DR, Schneider TW, Montgomery CB, Custer TL, Schaffner AE, Liu QY, Li YX, Barker JL, Hickman JJ. Developmental neurobiology implications from fabrication and analysis of hippocampal neuronal networks on patterned silane- modified surfaces. J Am Chem Soc 1998;120(47): 12169–12177.
19. Plueddemann EP. Silane adhesion promoters in coatings. Progr Organic Coatings 1983;11(3):297–308.
20. Whitesides GM, Laibinis P, Folkers J, Prime K, Seto C, Zerkowski J. Self-assembly—alkanethiolates on gold and hydrogen-bonded networks. Abstr Papers Am Chem Soc 1991;201:103-INOR.
21. Gillen G, Wight S, Bennett J, Tarlov MJ. Patterning of self-assembled alkanethiol monolayers on silver by microfocus ion and electron-beam bombardment. Appl Phys Lett 1994;65(5): 534–536.
22. Stenger DA, Hickman JJ, Bateman KE, Ravenscroft MS, Ma W, Pancrazio JJ, Shaffer K, Schaffner AE, Cribbs DH, Cotman CW. Microlithographic determination of axonal/dendritic polarity in cultured hippocampal neurons. J Neurosci Methods 1998;82(2):167–173.
23. Matsuda T, Sugawara T. Development of surface photochemical modification method for micropatterning of cultured-cells. J Biomed Mater Res 1995;29(6):749–756.
24. Dulcey CS, Georger JH, Chen MS, McElvany SW, Oferrall CE, Benezra VI, Calvert JM. Photochemistry and patterning of self-assembled monolayer films containing aromatic hydrocarbon functional groups. Langmuir 1996;12(6):1638–1650.
25. Romanova EV, Fosser KA, Stanislav SR, Nuzzo RG, Sweedler JV. Engineering the morphology and electrophysiological parameters of cultured neurons by microfluidic surface patterning. FASEB J 2004.
26. Corey JM, Wheeler BC, Brewer GJ. Compliance of hippocampal neurons to patterned substrate networks. J Neurosci Res 1991;30(2):300–307.
27. Griscom L, Degenaar P, LePioufle B, Tamiya E, Fujita H. Techniques for patterning and guidance of primary culture neurons on micro-electrode arrays. Sen Actuators B 2002;83 (1–3):15–21.
28. Nicolau DV, Taguchi T, Taniguchi H, Tanigawa H, Yoshikawa S. Patterning neuronal and glia cells on light-assisted functionalised photoresists. Biosens Bioelectron 1999;14(3):317–325.
29. He W, Halberstadt CR, Gonsalves KE. Lithography application of a novel photoresist for patterning of cells. Biomaterials 2004;11:2055–8063.
30. Liu GY, Amro NA. Positioning protein molecules on surfaces: A nanoengineering approach to supramolecular chemistry. Proc Nat Acad Sci 2002;99(8):5165–5170.
31. Pirrung MC, Huang CY. A general method for the spatially defined immobilization of biomolecules on glass surfaces using "caged" biotin. Bioconjugate Chem 1996;7(3):317–321.
32. Sigrist H, Collioud A, Clemence JF, Gao H, Luginbuhl R, Sanger M, Sundarababu G. Surface immobilization of biomolecules by light. Opt Eng 1995;34(8):2339–2348.
33. Herbert CB, McLernon TL, Hypolite CL, Adams DN, Pikus L, Huang CC, Fields GB, Letourneau PC, Distefano MD, Hu WS. Micropatterning gradients and controlling surface densities of photoactivatable biomolecules on self-assembled monolayers of oligo(ethylene glycol) alkanethiolates. Chem Bio 1997;4(10): 731–737.
34. Sorribas H, Padeste C, Tiefenauer L. Photolithographic generation of protein micropatterns for neuron culture applications. Biomaterials 2002;23(3):893–900.
35. Lee K-N, Shin D-S, Lee Y-S, Kim Y-K. Protein patterning by virtual mask photolithography using a micromirror array. J Micromech Microeng 2003;13(1):18–25.
36. Chen GP, Imanishi Y, Ito Y. Effect of protein and cell behavior on pattern-grafted thermoresponsive polymer. J Biomed Mater Res 1998;42(1):38–44.
37. Singhvi R, Kumar A, Lopez GP, Stephanopoulos GN, Wang DI, Whitesides GM, Ingber DE. Engineering cell shape and function. Science 1994;264(5159):696–698.
38. Lahiri J, Ostuni E, Whitesides GM. Patterning ligands on reactive SAMs by microcontact printing. Langmuir 1999;15 (6):2055–2060.
39. Cornish T, Branch DW, Wheeler BC, Campanelli JT. Microcontact printing: A versatile technique for the study of synaptogenic molecules. Mol Cell Neurosci 2002;20(1):140–153.
40. Kane RS, Takayama S, Ostuni E, Ingber DE, Whitesides GM. Patterning proteins and cells using soft lithography. Biomaterials 1999;20(23–24):2363–2376.
41. Oliva AA, James CD, Kingman CE, Craighead HG, Banker GA. Patterning axonal guidance molecules using a novel strategy for microcontact printing. Neurochem Res 2003;28 (11):1639–1648.
42. Martin BD, Gaber BP, Patterson CH, Turner DC. Direct protein microarray fabrication using a hydrogel "stamper". Langmuir 1998;14(15):3971–3975.
43. Mayer M, Yang J, Gitlin I, Gracias DH, Whitesides GM. Micropatterned agarose gels for stamping arrays of proteins and gradients of proteins. Proteomics 2004;4(8):2366–2376.
44. Lauer L, Ingebrandt S, Scholl M, Offenhausser A. Aligned microcontact printing of biomolecules on microelectronic device surfaces. IEEE Trans Biomed Eng 2001;48(7):838–842.
45. Craighead HG, James CD, Turner AMP. Chemical and topographical patterning for directed cell attachment. Curr Opin Solid State Mater Sci 2001;5(2–3):177–184.
46. Roth EA, Xu T, Das M, Gregory C, Hickman JJ, Boland T. Inkjet printing for high-throughput cell patterning. Biomaterials 2004;25(17):3707.
47. Newman JD, Turner APF, Marrazza G. Ink-jet printing for the fabrication of amperometric glucose biosensors. Anal Chim Acta 1992;262(1):13–17.
48. Xu T, Jin J, Gregory C, Hickman JJ, Boland T. Inkjet printing of viable mammalian cells. Biomaterials 2005;26(1):93–99.
49. Schena M, Heller RA, Thieriault TP, Konrad K, Lachenmeier E, Davis RW. Microarrays: Biotechnology's discovery platform for functional genomics. Trends Biotechnol 1998;16(7):301–306.
50. Roda A, Guardigli M, Russo C, Pasini P, Baraldini M. Protein microdeposition using a conventional ink-jet printer. Biotechniques 2000;28(3):492–496.
51. Flaim CJ, Chien S, Bhatia SN. An extracellular matrix microarray for probing cellular differentiation. Nature Methods 2005;2(2):119–125.
52. Folch A, Toner M. Cellular micropatterns on biocompatible materials. Biotechnol Progr 1998;14(3):388–392.
53. Chiu DT, Jeon NL, Huang S, Kane RS, Wargo CJ, Choi IS, Ingber DE. Whitesides GM. Patterned deposition of cells and proteins onto surfaces by using three-dimensional microfluidic systems. Proc Natl Acad Sci USA 2000;97(6):2408–2413.
54. Takayama S, McDonald JC, Ostuni E, Liang MN, Kenis PJA, Ismagilov RF, Whitesides GM. Patterning cells and their

environments using multiple laminar fluid flows in capillary networks. Proc Natl Acad Sci USA 1999;96(10):5545–5548.
55. Delamarche E, Bernard A, Schmid H, Michel B, Biebuyck H. Patterned delivery of immunoglobulins to surfaces using microfluidic networks. Science 1997;276(5313):779–781.
56. Itoga K, Yamamoto JK, Kikuchi A, Okano T. Micropatterned surfaces prepared using a liquid crystal projector-modified photopolymerization device and microfluidics. J Biomed Mater Res 2004;69A:391–397.
57. Tan W, Desai TA. Microscale multilayer cocultures for biomimetic blood vessels. J Biomed Mater Res 2005;72A(2):146–160.
58. Tan W, Desai TA. Layer-by-layer microfluidics for biomimetic three-dimensional structures. Biomaterials 2004;25(7–8): 1355–1364.
59. Curtis A, Wilkinson C. Topographical control of cells. Biomaterials 1997;18(24):1573–1583.
60. Turner S, Kam L, Isaacson M, Craighead HG, Shain W, Turner J. Cell attachment on silicon nanostructures. J Vacuum Sci Technol B 1997;15(6):2848–2854.
61. Mahoney MJ, Chen RR, Tan J, Saltzman WM. The influence of microchannels on neurite growth and architecture. Biomaterials 2005;26(7):771–778.
62. Tan JL, Tien J, Pirone DM, Gray DS, Bhadriraju K, Chen CS. Cells lying on a bed of microneedles: An approach to isolate mechanical force. Proc Natl Acad Sci USA 2003;100(4):1484–1489.
63. Recknor JB, Recknor JC, Sakaguchi DS, Mallapragadaa SK. Oriented astroglial cell growth on micropatterned polystyrene substrates. Biomaterials 2004;25(14):2753–2767.
64. Xi JZ, Schmidt J, Montemagno C. Development of self-assembled muscle-MEMS microdevices. Biophys J 2004;86 (1):481A.
65. Albrecht DR, Tsang VL, Sah RL, Bhatia SN. Photo- and electropatterning of hydrogel-encapsulated living cell arrays. Lab Chip 2005;5(1):111–118.
66. Bloch J, Fine EG, Bouche N, Zurn AD, Aebischer P. Nerve growth factor- and neurotrophin-3-releasing guidance channels promote regeneration of the transected rat dorsal root. Exper Neurol 2001;172(2):425–432.
67. Folch A, Jo BH, Hurtado O, Beebe DJ, Toner M. Microfabricated elastomeric stencils for micropatterning cell cultures. J Biomed Mater Res 2000;52(2):346–353.
68. Ostuni E, Kane R, Chen CS, Ingber DE, Whitesides GM. Patterning mammalian cells using elastomeric membranes. Langmuir 2000;16(20):7811–7819.
69. Gole A, Sastry M. A new method for the generation of patterned protein films by encapsulation in arrays of thermally evaporated lipids. Biotechnol Bioeng 2001;74(2):172–178.
70. Shiku H, Uchida I, Matsue T. Microfabrication of alkylsilanized glass substrate by electrogenerated hydroxyl radical using scanning electrochemical microscopy. Langmuir 1997;13(26):7239–7244.
71. Turyan I, Matsue T, Mandler D. Patterning and characterization of surfaces with organic and biological molecules by the scanning electrochemical microscope. Anal Chem 2000;72 (15):3431–3435.
72. Amro NA, Xu S, Liu GY. Patterning surfaces using tip-directed displacement and self-assembly. Langmuir 2000; 16(7): 3006–3009.
73. Schwartz PV. Molecular transport from an atomic force microscope tip: A comparative study of dip-pen nanolithography. Langmuir 2002;18(10):4041–4046.
74. Agarwal G, Sowards LA, Naik RR, Stone MO. Dip-pen nanolithography in tapping mode. J Am Chem Soc 2003; 125(2):580–583.
75. Chang JC, Brewer GJ, Wheeler BC. A modified microstamping technique enhances polylysine transfer and neuronal cell patterning. Biomaterials 2003;24:2862–2870.
76. Curtis A, Riehle M. Tissue engineering: the biophysical background. Phys Med Biol 2001; 46(4):R47–R65.
77. Curtis A, Wilkinson C. Reactions of cells to topography. J Biomater Sci Pol Educ 1998;9:1313–1329.
78. Li B, Ma Y, Wang S, Moran PM. A technique for preparing protein gradients on polymeric surfaces: effects on PC12 pheochromocytoma cells. Biomaterials 2005;26:1487–1495.
79. McFarland CD, Thomas CH, DeFilippis C, Stelle JG, Healy KE. Protein adsorption and cell attachment to patterned surfaces. J Biomed Mater Res 2000;49(2):200–210.
80. Veiseh M, Wickes BT, Castner DG, Zhang MQ. Guided cell patterning on gold-silicon dioxide substrates by surface molecular engineering. Biomaterials 2004;25(16):3315–3324.
81. Zheng H, Berg MC, Rubner MF, Hammond PT. Controlling cell attachment selectively onto biological polymer-colloid templates using polymer-on-polymer stamping. Langmuir 2004;20(17):7215–7222.
82. Ber S, Kose GT, Hasirci V. Bone tissue engineering on patterned collagen films: An *in vitro* study. Biomaterials 2005; 16:1977–1986.
83. Dowell-Mesfin NM, Abdul-Karim MA, Turner AM, Schanz S, Craighead HG, Roysam B, Turner JN, Shain W. Topographically modified surfaces affect orientation and growth of hippocampal neurons. J Neural Eng 2004;1(2): 78–90.
84. Mohammed JS, DeCoster MA, McShane MJ. Micropatterning of nanoengineered surfaces to study neuronal cell attachment *in vitro*. Biomacromolecules 2004;5(5):1745–1755.
85. Mrksich M, Chen CS, Xia YN, Dike LE, Ingber DE, Whitesides GM. Controlling cell attachment on contoured surfaces with self- assembled monolayers of alkanethiolates on gold. Proc Natl Acad Sci USA 1996;93(20):10775–10778.
86. Theibaud P, Lauer L, Knoll W, Offenhausser A. PDMS device for patterned application of microfluids to neuronal cells arranged by microcontact printing. Biosen Bioelectro 2002; 17:87–93.
87. Scholl M, Sprossler C, Denyer M, Krause M, Nakajima K, Maelicke A, Knoll W, Offenhausser A. Ordered networks of rat hippocampal neurons attached to silicon oxide surfaces. J Neurosci Methods 2000;104(1):65–75.
88. Deutsch J, Motiagh D, Russell B, Desai TA. Fabrication of microtextured membranes for cardiac myocyte attachment and orientation. J Biomed Mater Res 2000;53(3):267–275.
89. Goessl A, Bowen-Pope DF, Hoffman AS. Control of shape and size of vascular smooth muscle cells *in vitro* by plasma lithography. J Biomed Mater Res 2001;57(1):15–24.
90. Motlagh D, Senyo SE, Desai TA, Russell B. Microtextured substrata alter gene expression, protein localization and the shape of cardiac myocytes. Biomaterials 2003;24(14):2463–2476.
91. Zinger O, Zhao G, Schwartz Z, Simpson J, Wieland M, Landolt D, Boyan B. Differential regulation of osteoblasts by substrate microstructural features. Biomaterials 2005;26 (14):1837–1847.
92. Matsuzawa M, Liesi P, Knoll W. Chemically modifying glass surfaces to study substratum-guided neurite outgrowth in culture. J Neurosci Methods 1996;69(2):189–196.
93. Clark P, Britland S, Connolly P. Growth cone guidance and neuron morphology on micropatterned laminin surfaces. J Cell Sci 1993;105:203–212.

94. Saneinejad S, Shoichet MS. Patterned poly(chlorotrifluoroethylene) guides primary nerve cell adhesion and neurite outgrowth. J Biomed Mater Res 2000;50(4):465–474.
95. Tai H, Buettner HM. Neurite outgrowth and growth cone morphology on micropatterned surfaces. Biotechnol Prog 1998;14:364–370.
96. Zhang ZP, Yoo R, Wells M, Beebe TP, Biran R, Tresco P. Neurite outgrowth on well-characterized surfaces: Preparation and characterization of chemically and spatially controlled fibronectin and RGD substrates with good bioactivity. Biomaterials 2005;26(1):47–61.
97. Heller DA, Garga V, Kelleher KJ, Lee TC, Mahbubani S, Sigworth LA, Lee TR, Rea MA. Patterned networks of mouse hippocampal neurons on peptide-coated gold surfaces. Biomaterials 2005;26(8):883–889.

See also BIOCOMPATIBLITY OF MATERIALS; BIOMATERIALS, SURFACE PROPERTIES OF; BIOMATERIALS: TISSUE ENGINEERING AND SCAFFOLDS.

BIOMEDICAL EQUIPMENT MAINTENANCE.

See EQUIPMENT MAINTENANCE, BIOMEDICAL.

BIOSENSORS.

See IMMUNOLOGICALLY SENSITIVE FIELD-EFFECT TRANSISTORS.

BIOTELEMETRY

BABAK ZIAIE
Purdue University
W. Lafayette, Indiana

INTRODUCTION

The ability to use wireless techniques for measurement and control of various physiological parameters inside human and animal bodies has been a long-term goal of physicians and biologists going back to the early days of wireless communication. From early on, it was recognized that this capability could provide effective diagnostic, therapeutic, and prosthetic tools in physiological research and pathological intervention. However, this goal eluded scientists prior to the invention of transistor in 1947. Vacuum tubes were too bulky and power hungry to be of any use in many wireless biomedical applications. During the late 1950s, MacKay performed his early pioneering work on what he called Endoradiosonde (1). This was a single-transistor blocking oscillator designed to be swallowed by a subject and was able to measure pressure and temperature in the digestive track. Following this early work, came a number of other simple discrete systems each designed to measure a specific parameter (temperature, pressure, force, flow, etc.) (2). By the late 1960s, progress in the design and fabrication of integrated circuits provided an opportunity to expand the functionality of these early systems. Various hybrid single and multichannel telemetry systems were developed during the 1970s and the 1980s (3). In addition, implantable therapeutic and prosthetic devices started to appear in the market. Cardiac pacemakers and cochlear prosthetics proved effective and reliable enough to be implanted in thousands of patients. We direct the interested readers to several excellent reviews published over the past several decades summarizing these advances in their perspective time periods. These include a review article by W. H. Ko and M. R. Neuman in the *Science* covering the technologies available in the 1960s (4) and another similar paper by Topich covering the 1970s period (5). Three subsequent reviews detailed the efforts in the 1980s (6–8) followed by the most recent article published in 1999 (9). An outdated, but classic reference book in biotelemetry, is by MacKay, which still can be used as a good starting point for some simple single channel systems and includes some ingenious techniques used by early investigators to gain remote physiological information (10).

The latter part of the 1990s witnessed impressive advances in microelectromechanical (MEMS) based transducer and packaging technology, new and compact power sources (high efficiency inductive powering and miniature batteries), and CMOS low power wireless integrated circuits that provided another major impetus to the development of biotelemetry systems (11–18). These advances have created new opportunities for increased reliability and functionality, which had been hard to achieve with pervious technologies. The term biotelemetry itself has been for most part superseded by Microbiotelemetry or Wireless Microsystems to denote these recent changes in technology. Furthermore, the burgeoning area of nanotechnology is poised to further enhance these capabilities beyond what have been achievable using current miniaturization techniques. This is particularly true in the biochemical sensing and chemical delivery areas and will undoubtedly have a major impact on the future generations of implantable biotelemetry microsystems.

This review article is intended to complement and expand the earlier reviews by emphasizing newer developments in the area of biomedical telemetry in particular attention is paid to the opportunities created by recent advances in the area of microbiotelemetry (i.e., systems having volumes $\sim$1 cm^3 or less) by low power CMOS wireless integrated circuits, micromachined–MEMS transducers, biocompatible coatings, and advanced batch-scale packaging. We have both expanded and narrowed the traditional definition of biotelemetry by including therapeutic–rehabilitative microsystems and excluding wired devices that although fit under the strict definition of biotelemetry; do not constitute an emerging technology. In the following sections, after discussing several major components of such biotelemetry microsystems, such as transducers, interface electronics, wireless communication, power sources, and packaging, we will present some selected examples to demonstrate the state of the art. These include implantable systems for biochemical and physiological measurements, drug delivery microsystems, and neuromuscular and visual prosthetic devices. Although our primary definition of biotelemetry encompass devices with active electronics and signal processing capabilities, we will also discuss passive MEMS-based transponders that do not require on-board signal processing and can be interrogated using simple radio-frequency (rf) techniques. Finally, we should mention that although in a strict sense biotelemetry encompasses systems targeted for physiological measurements, this

narrow definition is no longer valid or desirable. A broader scope including neuromuscular stimulation and chemical delivery is currently understood to be more indicative of the term biotelemetry.

BIOTELEMETRY SYSTEMS

For the purpose of current discussion biotelemetry systems can be defined as a group of medical devices that (*1*) incorporate one or several miniature transducers (i.e., sensors and actuators), (*2*) have an on-board power supply (i.e., battery) or are powered from outside using inductive coupling, (*3*) can communicate with outside (bidirectional or unidirectional) through an rf interface, (*4*) have on-board signal processing capability, (*5*) are constructed using biocompatible materials, and (*6*) use advanced batch-scale packaging techniques. Although one microsystem might incorporate all of the above components, the demarcation line is rather fluid and can be more broadly interpreted. For example, passive MEMS-based microtransponders do not contain on-board signal processing capability, but use advanced MEMS packaging and transducer technology and are usually considered to be a telemetry device. We should also emphasize that the above components are interrelated and a good system designer must pay considerable attention from the onset to this fact. For example, one might have to choose a certain power source or packaging scheme to accommodate the desired transducer, interface electronics, and wireless communication. In the following sections, we will discuss various components of a typical biotelemetry system with more attention being paid to the wireless communication block. For other components, we provide a brief discussion highlighting major recent developments and refer the reader to some recent literature is these areas.

Transducers

Transducers are interfaces between biological tissue and readout electronics–signal processing. Their performance is critical to the success of the overall microsystem (19–24). Current trend in miniaturization of transducers and their integration with signal processing circuitry have considerably enhanced their performance. This is particularly true with respect to MEMS-based sensors and actuators, where the advantages of miniaturization have been prominent. Development in the area of microactuators has been lagging behind the microsensors due to the inherent difficulty in designing microdevices that efficiently and reliably generate motion. Although some transducing schemes, such as electrostatic force generation, has advantageous scaling properties in the microdomain, problems associated with packaging and reliability has prevented their successful application. The MEMS-based microsensors have been more successful and offer several advantages compared to the macrodomain counterparts. These include lower power consumption, increased sensitivity, higher reliability, and lower cost due to batch fabrication. However, they suffer from a poor signal/noise ratio, hence requiring a close by interface circuit. Among the many microsensors designed and fabricated over the past two decades, physical sensors have been by and large more successful. This is due to their inherent robustness and isolation from any direct contact with biological tissue in sensors, such as accelerometers and gyroscopes. Issues related to packaging and long-term stability have plagued the implantable chemical sensors. Long-term baseline and sensitivity stability are major problems associated with implantable sensors. Depending on the type of the sensor, several different factors contribute to the drift. For example, in implantable pressure sensors, packaging generated stresses due to thermal mismatch and long-term material creep are the main sources of baseline drift. In chemical sensors, biofouling and fibrous capsule formation is the main culprit. Some of these can be mitigated through clever mechanical design and appropriate choice of material, however, some are more difficult to prevent (e.g., biofouling and fibrous capsule formation). Recent developments in the area of antifouling material and controlled release have provided new opportunities to solve some of these long standing problems (25–27).

Interface Electronics

As mentioned previously, most miniature and MEMS-based transducers suffer from poor signal/noise ratio and require on-board interface electronics. This, of course, is also more essential for implantable microsystems. The choice of integrating the signal processing with the MEMS transducer on the same substrate or having a separate signal processing chip in close proximity depends on many factors, such as process complexity, yield, fabrication costs, packaging, and general design philosophy. Except for post-CMOS MEMS processing methods, which rely on undercutting micromechanical structures subsequent to the fabrication of the circuitry (28), other integrated approaches require extensive modifications to the standard CMOS processes and have not been able to attract much attention. Post-CMOS processing is an attractive approach although packaging issues still can pose roadblocks to successful implementation. Hybrid approach has been typically more popular with the implantable biotelemetry microsystem designers providing flexibility at a lower cost. Power consumption is a major design consideration in implantable wireless microsystems that rely on batteries for an energy source. Low power and subthreshold CMOS design can reduce the power consumption to nanowatt levels (29–33). Important analogue and mixed-signal building blocks for implantable wireless microsystems include amplifiers, oscillators, multiplexers, A/D and D/A converters, and voltage references. In addition, many such systems require some digital signal processing and logic function in the form of finite-state machines. In order to reduce the power consumption, it is preferable to perform the DSP functions outside the body although small finite-state machines can be implemented at low power consumptions.

Wireless Communication

The choice of appropriate communication scheme for a biotelemetry system depends on several factors, such as (*1*) number of channels, (*2*) device lifetime, and (*3*) transmission range. For single (or two) channel systems, one can choose a variety of modulation schemes and techniques.

These systems are the oldest type of biotelemetry devices (1) and can range from simple blocking oscillators to single channel frequency modulation (FM) transmitters. They are attractive since one can design a prototype rather quickly using off-the-shelf components. Figure 1 shows a schematic of the famous blocking oscillator first used by MacKay to transmit pressure and temperature (10). It consists of a single bipolar transistor oscillator configured to periodically turn itself on and off. The oscillation frequency depends on the resonant frequency of the tank circuit that can be made to vary with parameters, such as pressure, by including a capacitive or inductive pressure sensor. The on–off repetition frequency can be made to depend on the temperature by incorporating a thermistor in the circuit. This is an interesting example of an ingenious design that can be accomplished with a minimum amount of effort and hardware. An example of a more recent attempt at single channel telemetry is a two-channel system designed by Mohseni et al. to transmit moth electromyograms (34). The circuit schematic and a picture of the fully assembled device are shown in Fig. 2. As can be seen, each channel

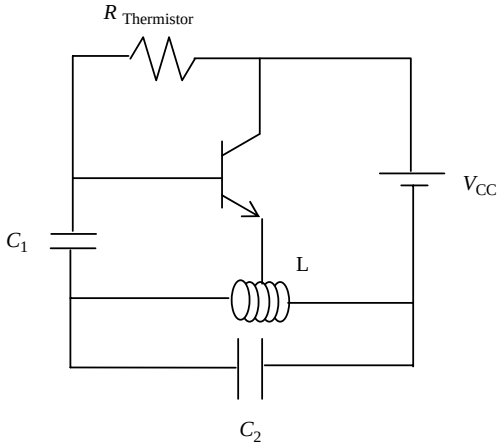
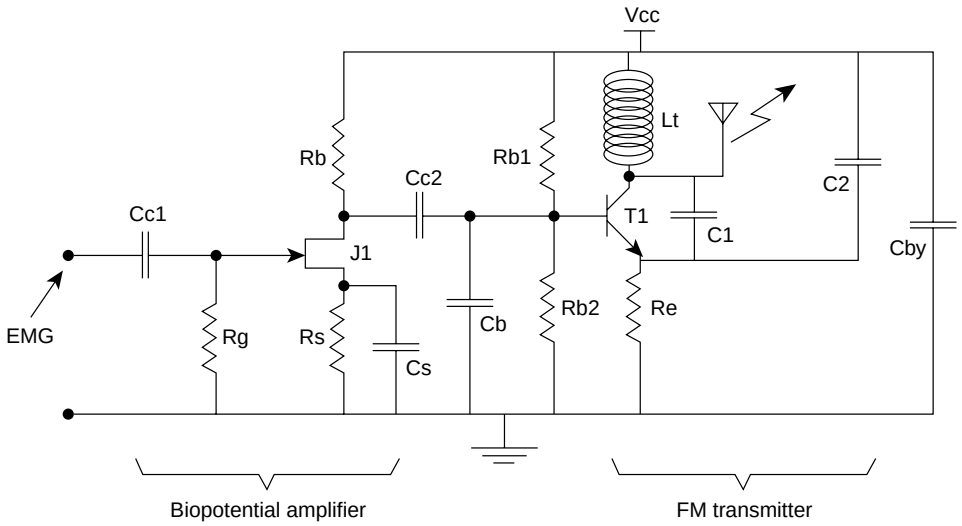


Figure 1. Schematic circuit of a blocking oscillator used to transmit pressure and temperature.



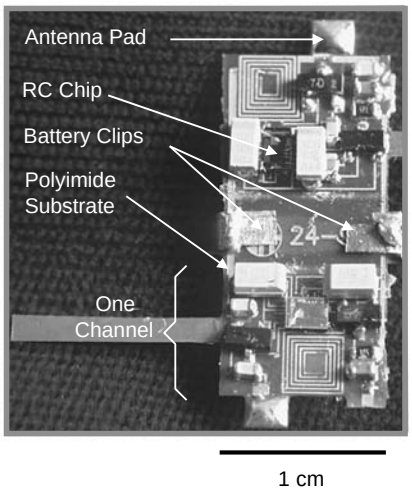


Figure 2. Schematic diagram and photograph of a biotelemetry system used to transmit flight muscle electromyograms in moths showing the polyimide flex circuit and various components (the Colpitts Oscillator inductor is used as the transmitting antenna).

consists of a biopotential amplifier followed by a Colpitts oscillator with operating frequency tunable in the 88–108 MHz commercial FM band. The substrate for the biotelemetry module was a polyimide flex circuit in order to reduce the weight such that the Moth can carry the system during flight. The overall system measures $10 \times 10 \times 3$ mm, weighs 0.74 g, uses two 1.5 V batteries, dissipates $\sim$ 2 mW, and has a transmission range of 2 m.

Multichannel systems are of more scientific and clinical interest. These systems rely on different and more elaborate communication schemes. For the purpose of current discussion, we will divide these systems into the ones that operate with a battery and the ones that are powered from outside using an inductive link. Battery-operated biotelemetry microsystems rely on different communication schemes than the inductively powered ones. Figure 3 shows a schematic block diagram of a time-division multiplexed multichannel system. It consists of several transducers with their associated signal conditioning circuits. These might include operations, such as simple buffering, low level amplification, filtering, or all three. Subsequent to signal conditioning, different channels are multiplexed using an analogue MUX. Although recent advances in AD technology might allow each channel to be digitized prior to multiplexing, this is not an attractive option for biotelemetry systems (unless there are only a few channels), since it requires an increase in power consumption that most biotelemetry systems cannot afford. All the timing and framing information is also added to the outgoing multiplexed signal at this stage. After multiplexing, an AD converter is used to digitize the signal. This is followed by a rf transmitter and a miniature antenna. The transmitted signal is picked up by a remote receiver and the signal is demodulated and separated accordingly. The described architecture is the one used currently by most investigators. Although over the years many different modulation scheme (pulse-width-modulation, pulse-position-modulation, pulse-amplitude-modulation, etc.) and system architectures have been tried; due to the proliferation of inexpensive integrated low power AD converters, the pulse-code-modulation (PCM) using an integrated AD is the dominant method these days.

The transmission of the digitized signal can be accomplished using any of the several digital modulation schemes (PAM, PFM, QPSK, etc.), which offer standard trade offs between transmitter and receiver circuit complexity, power consumption, and signal/noise ratio (35). Typical frequencies used in such systems are in the lower UHF range (100–500 MHz). Higher frequencies result in smaller transmitter antenna at the expense of increased tissue loss. Although tissue loss is a major concern in transmitting power to implantable microsystems, it is less of an issue in data transmission, since a sensitive receiver outside the body can easily demodulate the signal. Recent advances in low power CMOS rf circuit design has resulted in an explosive growth of custom made Application Specific Integrated Circuits (ASIC), and off-the-shelf rf circuits suitable for a variety of biotelemetry applications (36–38). In addition, explosive proliferation of wireless communication systems (cell phones, wireless PDAs, Wi-Fi systems, etc.) have provided a unique opportunity to piggyback major WLAN manufacturers and simplify the design of biotelemetry microdevices (39,40). This cannot only increase the performance of the system, but also creates a standard platform for many diverse applications. Although the commercially available wireless chips have large bandwidths and some superb functionality, their power consumption is higher than what is acceptable for many of the implantable microsystems. This, however, is going to change in the future by the aggressive move

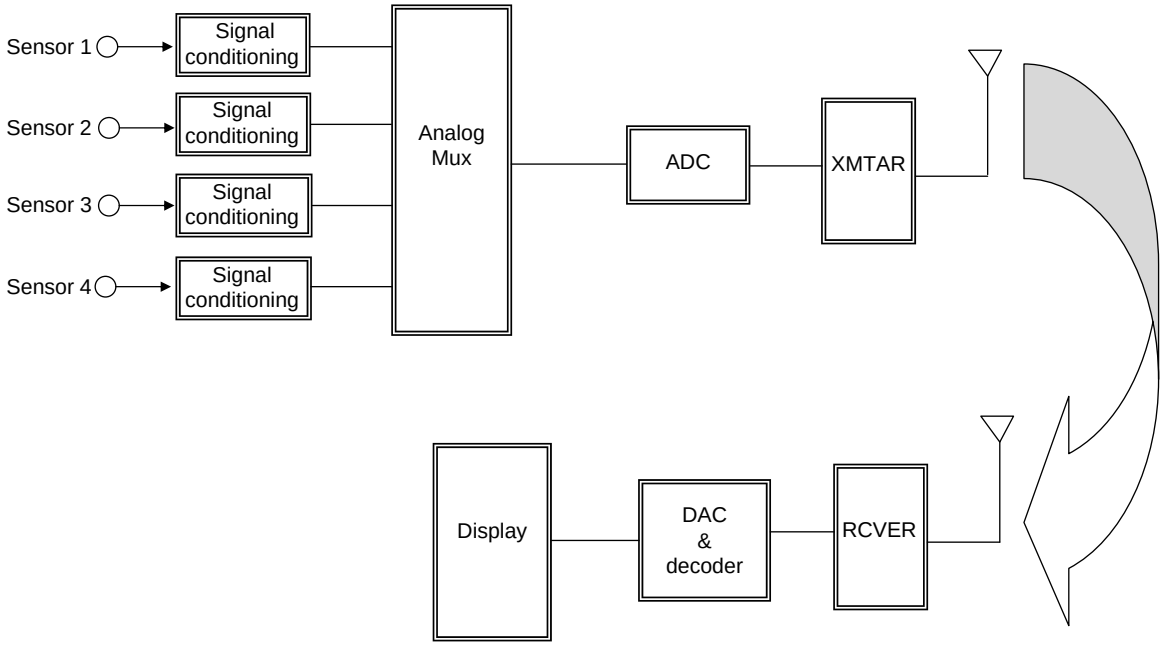


Figure 3. Block diagram of a multichannel biotelemetry system.

toward lower power handheld consumer electronics. A particularly attractive WLAN system suitable for biotelemetry is the Bluetooth system (41). This system, which was initially designed for wireless connection of multiple systems (computer, fax, printer, PDA, etc.) located in close proximity, has been adopted by many medical device manufacturer for their various biotelemetry applications. The advantage of Bluetooth compared to other Wi-Fi system, such as 902.11-b, is its lower power consumption at the expense of a smaller data rate (2.4 GHz carrier frequency, 1 Mbps data rate, and 10 m transmission range). This is not critical in most biotelemetry applications since the frequency bandwidth of most physiologically important signals are low (< 1 kHz). However, note that since the Bluetooth carrier frequency is rather high (2.4 GHz), the systems using Bluetooth or similar WLAN devices can not operate from inside the body and has to be worn by the subject on the outside.

Inductively powered telemetry systems differ from the battery operated ones in several important ways (42). First and foremost, the system has to be powered by an rf signal from outside; this puts several restrictions on frequency and physical range of operation. For implantable systems, the incoming signal frequency has to remain low in order for it to allow enough power to be coupled to the device (this means a frequency range of 1–10 MHz, see next section). In addition, if the device is small, due to a low coupling coefficient between the transmitter and receiver coil, the transmission range is usually limited to distances < 10 cm. Finally, in inductively powered systems, one has to devise a method to transmit the measured signal back to the outside unit. This can be done in several different ways with the load-modulation being the most popular method (43). In "load modulation", the outgoing digital stream of data is used to load the receiver antenna by switching a resistor in parallel with the tank circuit. This can be picked up through the transmitter coil located outside the body. A second technique that is more complex requires an on-chip transmitter and a second coil to transmit the recorded data at a different frequency. The inward link can be easily implemented using amplitude modulation, that is, the incoming rf signal that powers the microsystem is modulated by digitally varying the amplitude. It is evident that the modulation index cannot be 100% since that would cut off the power supply to the device (unless a storage capacitor is used). The coding scheme is based on the pulse time duration, that is,"1" and "0" have the same amplitude, but different durations (42). This modulation technique requires a simple detection circuitry (envelope detector) and is immune to amplitude variations, which are inevitable in such systems.

In addition to the above mentioned differences between the battery operated and inductively powered biotelemetry systems, the implanted circuit in the latter case also includes several modules that are unique and require special attention. These have mostly to do with power reception (rectifier and voltage regulator), clock extraction, and data demodulation. Figure 4 shows a block diagram of the receiver circuit for an inductively powered microsystem currently being developed in the author's laboratory for the measurement of intraocular pressure in glaucoma patients. It consists of a full-bridge rectifier, a voltage regulator, a piezoresistive pressure sensor, and voltage to frequency converter. The incoming rf signal is first rectified and used to generate a stable voltage reference being used by the rest of the circuit (amplifiers, filters, etc.). The clock is extracted from the incoming rf signal and is used wherever it is needed in the receiver circuit. The pressure sensor bridge voltage is first amplified and converted to a stream of pulses having a frequency proportional to the pressure. This signal is then used to load-modulate the tank circuit. The receiver circuitry for most of the reported inductively powered biotelemetry systems were fabricated through CMOS foundries, such as MOSIS. This is due to the fact that one can simply design a single chip performing all of the mentioned functions in a CMOS technology, and hence save valuable space. In the sections dealing with various applications, we will describe several other inductively powered telemetry systems.

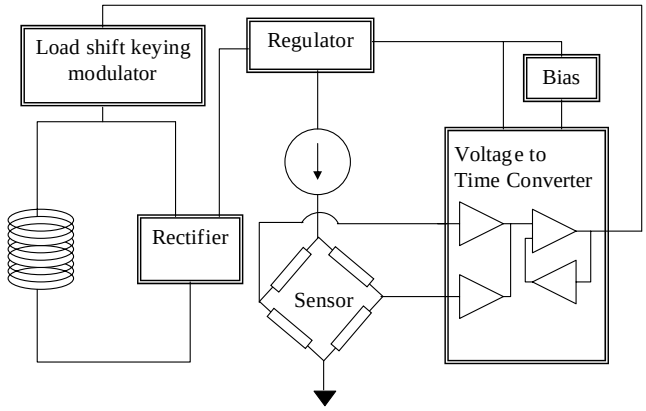


Figure 4. Block diagram of an implantable biotelemetry system used in the measurement of intraocular pressure in glaucoma patients.

There has not been much effort in the area of antenna design for biotelemetry applications. This is due to the basic fact that these systems are small and operate at low frequencies, hence, most antennas employed in such systems belong to the"small antenna" category, that is, the antenna size is much smaller than the wavelength. In such cases it is difficult to optimize the design and most investigators simply use a short electrical or magnetic dipole. For example, in many situations the inductor in the output stage can be used to transmit the information. Or alternatively, a short wire can be used in the transmitter as an electrical dipole. These antennas are usually low gain and have an omnidirectional pattern (44). Systems operating at higher frequencies, such as externally worn Wi-Fi modules, however, can benefit from an optimized design.

In addition to using an rf signal to transmit information that constitutes the majority of the work in the biotelemetry area, the use of ultrasound and infrared (IR) have also been explored by some investigators (45,46). The use of ultrasound is attractive in telemetering physiological information from aquatic animals and divers. This is due to the fact that rf signals are strongly absorbed by seawater while ultrasound is not affected to the same extent. The use of IR is also limited to some specific areas, such as systems that can be worn by the animal on the outside and are not

impeded by solid obstructions. This is due to the inability of IR to negotiate solid opaque objects (line of sight propagation) and its severe absorption by tissue. The advantage of free space IR transmission lies in its very wide bandwidth making it useful for transmitting neural signals. The rf, ultrasonic, and IR systems share many of the system components discussed so far, with the major difference between them having to do with the design and implementation of the output stage. The output transmitter for the ultrasonic biotelemetry systems is usually an ultrasonic transducer, such as PZT or PVDF, whereas for the IR systems it is usually a simple light-emitting diode (LED). The driver circuitry has to be able to accommodate the transducers, that is, a high voltage source for driving the ultrasonic element and a current.

Power Source

The choice of power source for implantable wireless microsystems depends on several factors, such as implant lifetime, system power consumption, temporal mode of operation (continuous or intermittent), and size. Progress in battery technology is incremental and usually several generations behind other electronic components (47). Although lithium batteries have been used in pacemakers for several years, they are usually large for microsystem applications. Other batteries used in hearing aids and calculators are smaller, but have limited capacity and can only be used for low power systems requiring limited lifespan or intermittent operation. Inductive powering is an attractive alternative for systems with large power requirements (e.g., neuromuscular stimulators) or long lifetime (e.g., prosthetic systems with >5 years lifetime) (14,15). In such systems, a transmitter coil is used to power a microchip using magnetic coupling. The choice of the transmission frequency is a trade-off between adequate miniaturization and tissue loss. For implantable microsystems, the frequency range of 1–10 MHz is usually considered optimum for providing adequate miniaturization while still staying below the high tissue absorption region ($>$10 MHz) (48). Although the link analysis and optimization methods have been around for many years (49), recent integration techniques that allow the fabrication of microcoils on top of CMOS receiver chip has allowed a new level of miniaturization (50). For applications that require the patient to carry the transmitter around, a high efficiency transmitter is needed in order to increase the battery lifetime. This is particularly critical in implantable microsystem, where the magnetic coupling between the transmitter and the receiver is low ($<1\%$). Class-E power amplifier/transmitters are popular among microsystem designers due to their high efficiency ($>80\%$) and relatively easy design and construction (51,52). They can also be easily amplitude modulated through supply switching.

Although ideally one would like to be able to tap into the chemical reservoir (i.e., glucose) available in the body to generate enough power for implantable microsystems (glucose-based fuel cell), difficulty in packaging and low efficiencies associated with such fuel cells have prevented their practical application (53). Thin-film batteries are also attractive, however, there still remain numerous material and integration difficulties that need to be resolved (54). Another alternative is nuclear batteries. Although they have been around for several decades and were used in some early pacemakers, safety and regulatory concerns forced medical device companies to abandon their efforts in this area. There has been a recent surge of interest in microsystem nuclear batteries for military applications (55). It is not hard to envision that due to the continuous decrease in chip power consumption and improve in batch scale MEMS packaging technology, one might be able to hermetically seal a small amount of radioactive source in order to power an implantable microsystem for a long period of time. Another possible power source is the mechanical movements associated with various organs. Several proposals dealing with parasitic power generation through tapping into this energy source have been suggested in the past few years (56). Although one can generate adequate power from activities, such as walking, to power an external electronic device, difficulty in efficient mechanical coupling to internal organ movements make an implantable device hard to design and utilize.

Packaging and Encapsulation

Proper packaging and encapsulation of biotelemetry microsystems is a challenging design aspect particularly if the device has to be implanted for a considerable period. The package must accomplish two tasks simultaneously: (*1*) protect the electronics from the harsh body environment while providing access windows for transducers to interact with the desired measurand, and (*2*) protect the body from possible hazardous material in the microsystem. The second task is easier to fulfill since there is a cornucopia of various biocompatible materials available to the implant designer (57). For example, silicon and glass, which are the material of choice in many MEMS applications, are both biocompatible. In addition, polydimethylsiloxane (PDMS) and several other polymers (e.g., polyimide, polycarbonate, parylene) commonly used in microsystem design are also accepted by the body. The first requirement is, however, more challenging. The degree of protection required for implantable microsystems depends on the required lifetime of the device. For short durations (several months), polymeric encapsulants might be adequate if one can conformally deposit them over the substrates (e.g., plasma deposited parylene) (58). These techniques are considered non-hermetic and have a limited lifetime. For long-term operation, hermetic sealing techniques are required (59). Although pacemaker and defibrillator industries have been very successful in sealing their systems in tight titanium enclosures; these techniques are not suitable for microsystem applications. For example a metallic enclosure prevents the transmission of power and data to the microsystem. In addition, these sealing methods are serial in nature (e.g., laser or electron beam welding) and are not compatible with integrated batch fabrication methods used in microsystem design. Silicon–glass electrostatic and silicon–silicon fusion bonding are attractive methods for packaging implantable microsystems (60). Both of these bonding methods are hermetic and can be performed at the wafer level. These are particularly attractive for

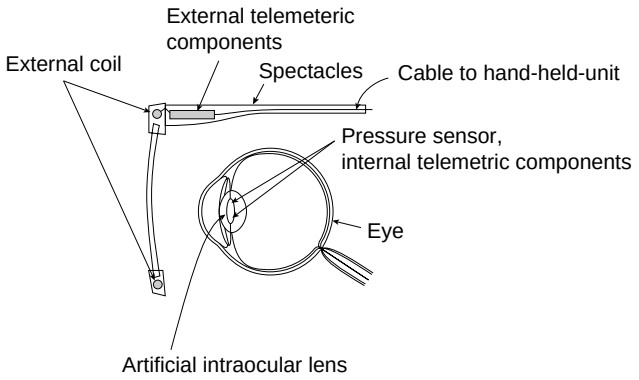


Figure 5. Schematic of the IOP measurement microsystem (61).

inductively powered wireless microsystems since most batteries cannot tolerate the high temperatures required in such substrate bondings. Other methods, such as metal electroplating, have also been used to seal integrated MEMS microsystems. However, their long-term performance is usually inferior to the anodic and fusion bondings. In addition to providing a hermetic seal, the package must allow feeedthrough for transducers located outside the package (18). In macrodevices, such as pacemakers, where the feedthrough lines are large and not too many, traditional methods, such as glass–metal or ceramic–metal has been employed for many years. In microsystems, such methods are not applicable and batch scale techniques must be adopted.

DIAGNOSTIC APPLICATIONS

Diagnostic biotelemetry microsystems are used to gather physiological or histological information from within the body in order to identify pathology. Two recent examples are discussed in this category. The first is a microsystem designed to be implanted in the eye and to measure the intraocular pressure in order to diagnose low tension glaucoma. The second system, although not strictly implanted, is an endoscopic wireless camera-pill designed to be swallowed in order to capture images from the digestive track.

Figure 5 shows the schematic diagram of the intraocular pressure (IOP) measurement microsystem (61,62). This device is used to monitor the IOP in patients suffering from low tension glaucoma, that is, the pressure measured in the doctor's office is not elevated (normal IOP is ~10–20 mmHg, 1.33–2.66 kPa) while the patient is showing optic nerve degeneration associated with glaucoma. There is great interest in measuring the IOP in such patients during their normal course of daily activity (exercising, sleeping, etc). This can only be achieved using a wireless microsystem. The system shown in Fig. 5 consists of an external transmitter mounted on a spectacle, which is used to power a microchip implanted in the eye. A surface micromachined capacitive pressure sensor integrated with CMOS interface circuit is connected to the receiving antenna. The receiver chip implemented in an n-well 1.2 μm CMOS technology has overall dimensions of 2.5×2.5 mm^2 and consumes 210 μW (Fig. 6). The receiver polyimide-based antenna is, however, much

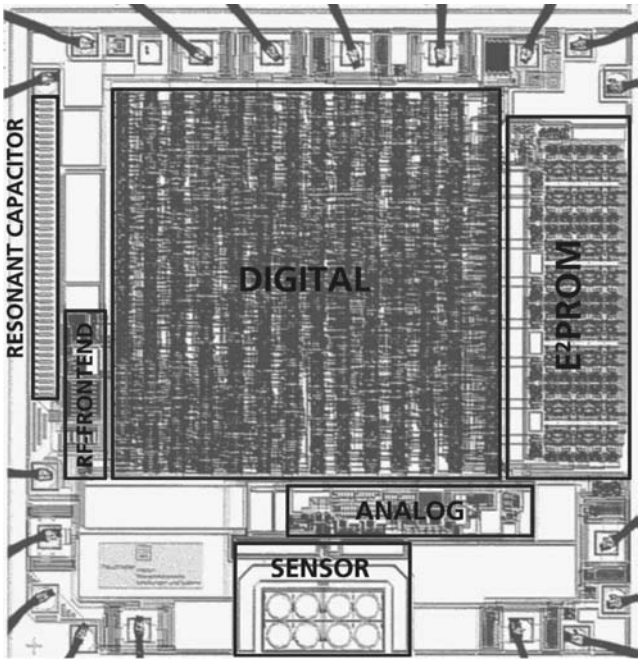


Figure 6. Micrograph of the IOP measurement microsystem receiver chip showing surface micromachined capacitive pressure sensors and other parts of the receiver circuitry (62).

larger (1 cm in diameter and connected to the receiver using flip chip bonding) requiring the device to be implanted along with an artificial lens. The incoming signal frequency is 6.78 MHz, while the IOP is transmitted at 13.56 MHz using load-modulation scheme. This example illustrates the levels of integration that can be achieved using low power CMOS technology, surface micromachining, and flip chip bonding.

The second example in the category of diagnostic microsystems is an endoscopic wireless pill shown in Fig. 7 (63,64). This pill is used to image small intestine, which is a particularly hard area to reach using current fiber optic technology. Although these days colonoscopy and gastroscopy are routinely performed, they cannot reach the small intestine and many disorders (e.g., frequent bleeding) in this organ have eluded direct examination. A wireless endoscopic pill cannot only image the small intestine, but also will reduce the pain and discomfort associated with regular gastrointestinal endoscopies. The endoscopic pill is a perfect example of what can be called *Reemerging Technology*, that is, the rebirth of an older technology based on new capabilities offered by advances in modern technology. Although the idea of a video pill is not new, before the development of low power microelectronics, white LED, CMOS image sensor, and wide-band wireless communication, fabrication of such a device was not feasible. The video pill currently marketed by Given Imaging Inc. is 11 mm in diameter and 30 mm in length (size of a large vitamin tablet) and incorporates: (*1*) a short focal length lens, (*2*) a CMOS image sensor (90,000 pixel), (*3*) four white LEDs, (*4*) a low power ASIC transmitter, and (*5*) two batteries (enough to allows the pill to go though the entire digestive track). The pill can capture and transmit

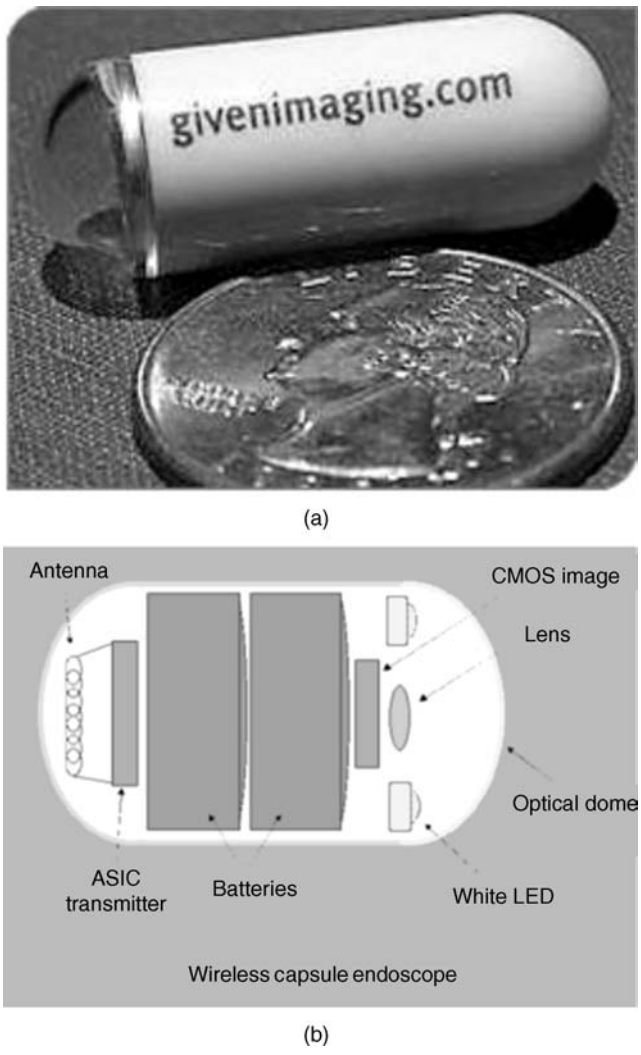


Figure 7. A photograph (a) and internal block diagram (b) of Given Imaging wireless endoscopic pill. (Courtesy Given Imaging.)

two images per second to an outside receiver capable of storing up to 5 h of data.

THERAPEUTIC APPLICATIONS

Therapeutic biotelemetry microsystems are designed to alleviate certain symptoms and help in the treatment of a disease. In this category, two such biotelemetry microsystems unit be described. The first is a drug delivery microchip designed to administer small quantities of potent drugs upon receiving a command signal from the outside. The second device is a passive micromachined glucose transponder, which can be used to remotely monitor glucose fluctuations allowing a tighter blood glucose control through frequent measurements and on-demand insulin delivery (pump therapy or multiple injections).

Figure 8 shows the central component of the drug delivery microchip (65,66). It consists of several microreservoirs (25 nL in volume) etched in a silicon substrate. Each microreservoir contains the targeted drug and is covered by a thin gold membrane (0.3 μm), which can be dissolved through the application of a small voltage (1 V vs. Saturated Calomel Electrode). The company marketing this technology (MicroCHIPS Inc.) is in the process of designing a wireless transceiver that can be used to address individual wells and release the drug upon the reception of the appropriate signal (67). Another company (ChipRx Inc.) is also aiming to develop a similar microsystem (Smart Pill) (68). Their release approach, however, is different and is based on conductive polymer actuators acting similar to a sphincter, opening and closing a tiny reservoir. Due to the potency of many drugs, safety and regulatory issues are more stringent in implantable drug delivery microsystems and will undoubtedly delay their appearance in the clinical settings.

Figure 9 shows the basic concept behind the glucose-sensitive microtransponder (69). A miniature MEMS-based microdevice is implanted in the subcutaneous tissue and an interrogating unit remotely measures the glucose levels without any hardwire connection. The microtrasponder is a passive LC resonator, which is coupled to a glucose-sensitive hydrogel. The glucose-dependent swelling and deswelling of the hydrogel is coupled to the resonator causing a change the capacitor value. This change translates into variations of the resonant frequency, which can be detected by the interrogating unit. Figure 10 shows the schematic drawing of the microtransponder with a capacitive sensing mechanism. The glucose sensitive hydrogel is mechanically coupled to a glass membrane and is separated from body fluids (in this case interstitial fluid) by a porous stiff plate. The porous plate allows the unhindered flow of water and glucose while blocking the hydrogel from escaping the cavity. A change in the glucose concentration of the external environment will cause a swelling or deswelling of the hydrogel, which will deflect the glass membrane and change the capacitance. The coil is totally embedded inside the silicon and can achieve a high quality factor and hence increased sensitivity by utilizing the whole wafer thickness (reducing the series resistance). The coil-embedded silicon and the glass substrate are hermetically sealed using glass–silicon anodic bonding.

REHABILITATIVE MICROSYSTEMS

Rehabilitative biotelemetry microsystems are used to substitute a lost function, such as vision, hearing, or motor activity. In this category, two microsystems are described. The first one is a single-channel neuromuscular microstimulator used to stimulate paralyzed muscle groups in paraplegic and quadriplegic patients. The second microsystem is a visual prosthetic device designed to stimulate ganglion cells in retina in order to restore vision to people afflicted with macular degeneration or retinitis pigmentosa.

Figure 11 shows a schematic of the single channel microstimulator (13). This device is $10 \times 2 \times 2$ mm^3 in dimensions and receives power and data through an inductively coupled link. It can be used to stimulate paralyzed muscle groups using thin-film microfabricated electrodes located at the ends of a silicon substrate. A hybrid capacitor

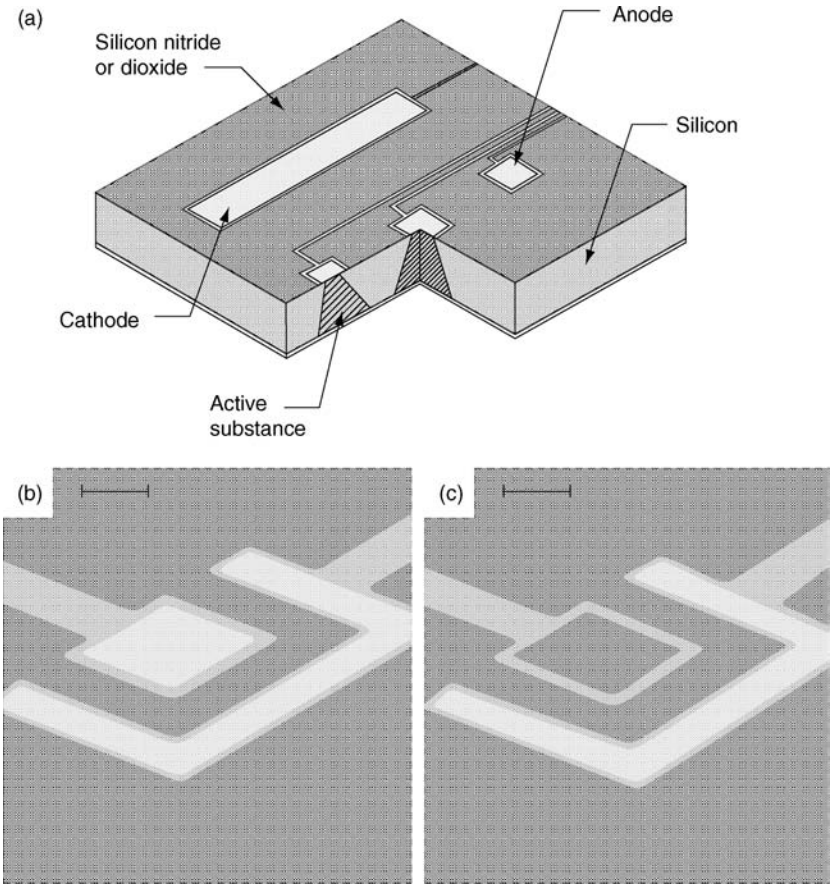


Figure 8. MicroCHIP drug delivery chip (a), a reservoir before and after dissolution of the gold membrane (b,c), the bar is 50 μm (65).

is used to store the charge in between the stimulation pulses and to deliver 10 mA of current to the muscle every 25 ms. A glass capsule hermetically seals a BiCMOS receiver circuitry along with various other passive components (receiver coil and charge storage capacitor) located on top of the silicon substrate. Figure 12 shows a photograph of the microstimulator in the bore of a gauge 10 hypodermic needle. As can be seen, the device requires a complicated hybrid assembly process in order to attach a wire-wound coil and a charge storage capacitor to the

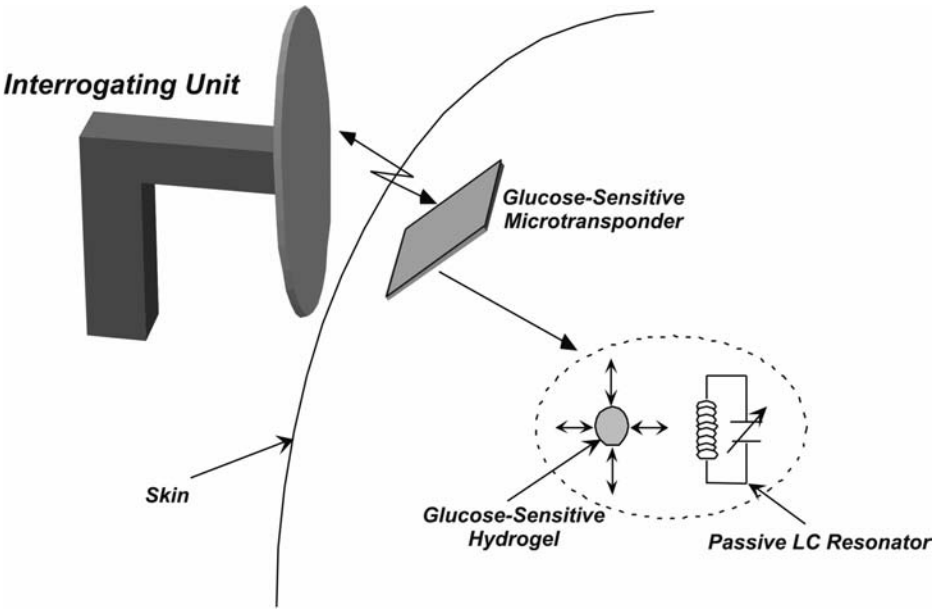


Figure 9. Basic concept behind the glucose-sensitive microtransponder.

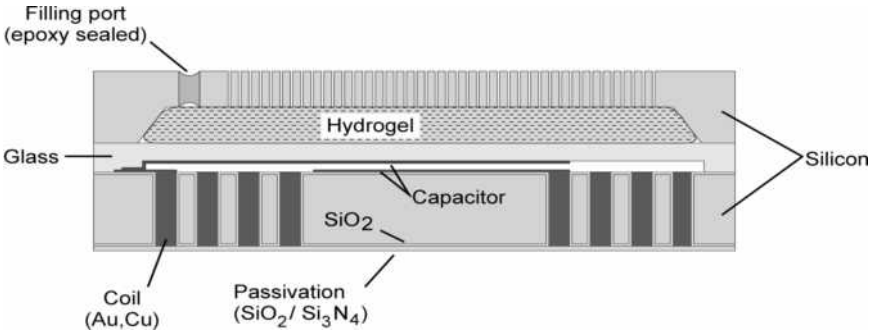


Figure 10. Cross-section of glucose microtransponder.

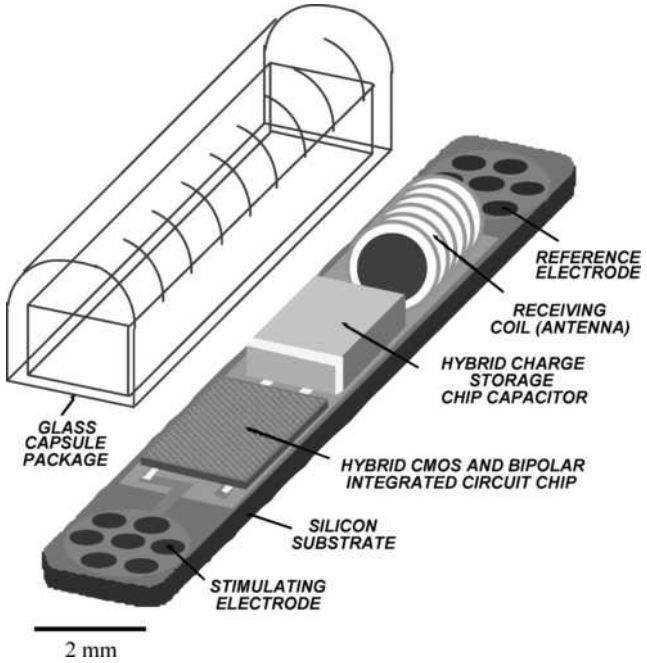


Figure 11. Schematic of a single-channel implantable neuromuscular microstimulator.

Figure 12. Photograph of the microstimulator in the bore of a gage 10 hypodermic needle.

receiver chip. In a subsequent design targeted for direct peripheral nerve stimulation (requiring smaller stimulation current), the coil was integrated on top of the BiCMOS electronics and on-chip charge storage capacitors were used thus considerably simplifying the packaging process. Figure 13 shows a micrograph of the chip with the electroplated copper inductor (70). A similar microdevice (i.e., a single channel microstimulator) was also developed by another group of investigators with the differences mainly related to the packaging technique (laser welding of a glass capsule instead of silicon–glass anodic bonding), chip technology (CMOS instead of BiCMOS), and electrode material (tantalum and iridium instead of iridium oxide) (42). Figure 14 shows a photograph of the microstimulator developed by Troyk, Loeb, and their colleagues.

Figure 15 shows the schematic of the visual prosthetic microsystem (71,72). A spectacle mounted camera is used to capture the visual information followed by digital conversion and transmission of data to a receiver chip

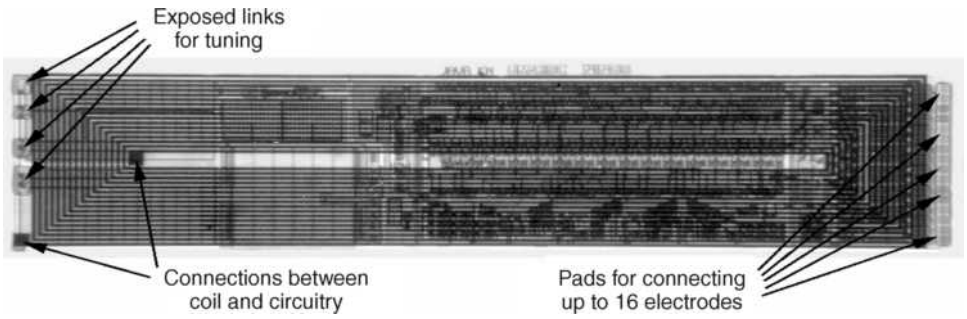


Figure 13. Microstimulator chip with integrated receiver coil and on-chip storage capacitor (70).

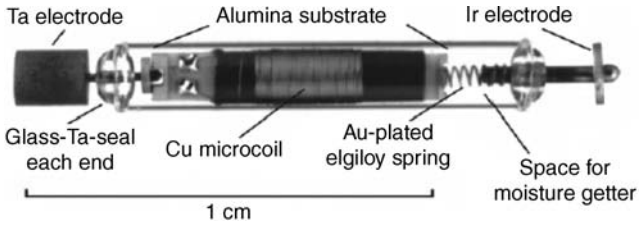


Figure 14. Photograph of a single channel microstimulator developed by Troyk (42).

implanted in the eye. The receiver uses this information to stimulate the ganglion cells in the retina through a microelectrode array in sub or epi-retinal location. This microsystem is designed for patients suffering from macular degeneration or retinitis pigmentosa. In both diseases, the light sensitive retinal cells (cones and rods) are destroyed while the more superficial retinal cells, that is, ganglion cells, are still viable and can be stimulated. Considering that macular degeneration is an age related pathology and will be afflicting more and more people as the average age of the population increases, such a microsystem will be of immense value in the coming decades. There are several groups pursuing such a device with different approaches to electrode placement (epi-or subretinal), chip design, and packaging. A German consortium that has also designed the IOP measurement microsystem is using a similar approach in antenna placement (receiver antenna in the lens), chip design, and packaging technology to implement a retinal prosthesis (61). Figure 16 shows photographs of the retinal stimulator receiver chip, stimulating electrodes, and polyimide antenna. The effort in the United States is moving along a similar approach (72,72).

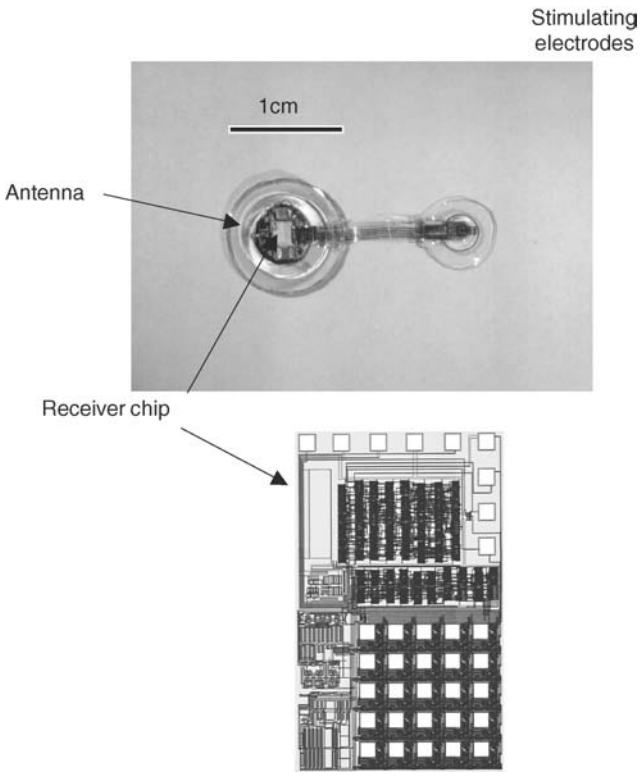


Figure 16. Retinal stimulator receiver chip, stimulating electrodes, and polyimide antenna (61). Chip size.

CONCLUSIONS

In this article, several biotelemetry microsystems currently being developed in the academia and industry were reviewed. Recent advances in MEMS-based transducers, low power CMOS integrated circuit, wireless communication transceivers, and advanced batch scale packaging have provided a unique opportunity to develop implantable biotelemetry microsystems with advanced functionalities not achievable previously. These systems will be indispensable to the twenty-first century physician by providing assistance in diagnosis and treatment. Future research and development will probably be focused on three areas: (*1*) nanotransducers, (*2*) self-assembly, and (*3*) advanced biomaterials. Although MEMS-based sensors and actuators have been successful in certain areas (particularly physical sensors), their performance could be further improved by utilizing nanoscale fabrication technology. This is particularly true in the area of chemical sensors where future diagnostic depends on detecting very small amounts of chemicals (usually biomarkers) well in advance of any

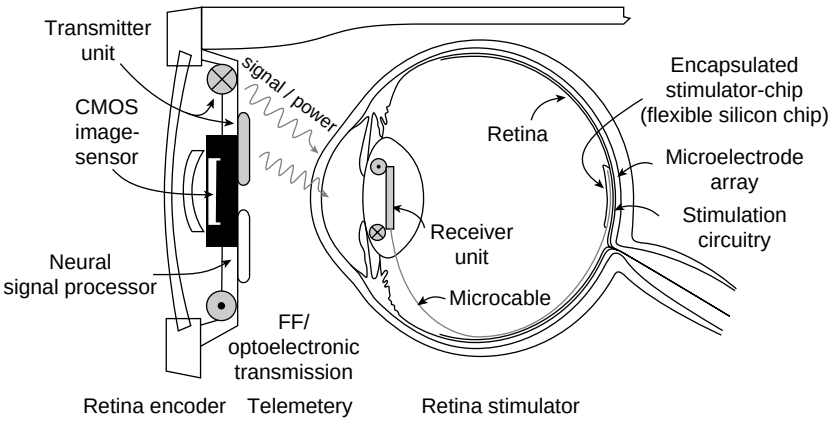


Figure 15. Schematic of a visual prosthetic microsystem (61).

physical sign. Nanosensors capable of high sensitivity chemical detection will be part of the future biotelemetry systems. In the actuator–delivery area, drug delivery via nanoparticles is a burgeoning area that will undoubtedly be incorporated into future therapeutic microsystems. Future packaging technology will probably incorporate self-assembly techniques currently being pursued by many micro–nanoresearch groups. This will be particularly important in microsystems incorporating multitude of nanosensors. Finally, advanced nanobased biomaterials will be used in implantable microsystems in order to enhance biocompatibility and prevent biofouling. These will include biocompatible surface engineering and interactive interface design (e.g., surfaces that release anti-inflammatory drugs in order to reduce postimplant fibrous capsule formation).

BIBLIOGRAPHY

Cited References

1. MacKay RS, Jacobson B. Endoradiosonde. Nature (London) 1957;179:1239–1240.
2. MacKay RS. Biomedical telemetry: The Formative Years. IEEE Eng Med Biol Mag 1983;2:11–17.
3. Knutti JW, Allen HV, Meindl JD. Integrated Circuit Implantable telemetry Systems. IEEE Eng Med Biol Mag 1983;2:47–50.
4. Ko WH, Neuman MR. Implant Biotelemetry and Microelectronics. Science 1867;156:351–360.
5. Topich JA. Medical Telemetry. CRC Handbook of Engineering in Medicine and Biology; 1976, p 41–75.
6. Jeutter DC. Biomedical Telemetry Techniques. CRC Crit Rev Biomed Eng 1982;11:121–174.
7. Meindl JD, et al. Implantable Telemetry. Methods Animal Exper 1986;3:37–111.
8. Kimmich HP. Biotelemetry. Encyclopedia of Medical Devices In: Webster JG, editor. 1988, p 409–425.
9. Santic A. Biomedical Telemetry. Wiley Encyclopedia of Electrical and Electronics Engineering, In: Webster JG, editor. 1999. p 438–454.
10. MacKay RS. Biomedical Telemetry, Sensing and Transmitting Biological Information from Animals and Man, 2nd ed. Piscataway (NJ): IEEE Press; 1993.
11. Wise KD. Special Issue on Sensors, Actuators, and Microsystems Proc. IEEE, 1998;86.
12. Cameron T, et al. Micromodular Implants to Provide Electrical Stimulation of Paralyzed Muscles and Limbs. IEEE Trans. Biomed. Eng. 1997;44:781–790.
13. Ziaie B, Nardin M, Coghlan AR, Najafi K. A Single Channel Microstimulator for Functional Neuromuscular Stimulation. IEEE Trans Biomed Eng 1997;44:909–920.
14. Hamici Z, Itti R, Champier J. A High-Efficiency Power and Data Transmission System for Biomedical Implanted Electronic Devices. Measurement Sci Technol. 1996;7:192–201.
15. Heetderks WJ. RF Powering of Millimeter- and Submillimeter-Sized Neural Prosthetic Implants. IEEE Trans Biomed Eng 1988;35:323–327.
16. Gray PR, Meyer RG. Future Directions in Silicon ICs for RF Personal Communications. Proceedings of the Custom Integrated Circuits Conference. 1995, p 83–89.
17. Abidi AA. RF CMOS Come of Age. IEEE Microwave Mag 2003;4:47–60.
18. Ziaie B, Von Arx JA, Dokmeci MR, Najafi K. A Hermetic Glass-Silicon Micropackage with High-Density on-chip Feedthroughs for Sensors and Actuators. IEEE J Microelectromech Systems, 1996;5:166–179.
19. Gopel W, Hesse J, Zemel JN. Sensors: A Comprehensive Survey, Vols. 1–8, New York: VCH Publishers; 1989.
20. Hohler JM, Sautz HP, editor. Microsystem Technology: A Powerful Tool for Biomolecular Studies. Boston: Birkhauser; 1999.
21. Taylor RF, Schultz JS. Handbook of Chemical and Biological Sensors, Boston: IOP Press; 1996.
22. Rogers EK. editor, Handbook of Biosensors and Electronic Nose, Boca Raton, (FL): CRC Press; 1997.
23. Hak GA. editor, The MEMS Handbook. Boca Raton (FL): CRC Press; 2001.
24. Webster JG. editor, The Measurement Instrumentation and Sensors Handbook, Boca Raton (FL): CRC Press; 1998.
25. Zhang M, Desai T, Ferrari M. Proteins and Cells on PEG Immobilized Silicon Surfaces. Biomaterials 1998;19:953–960.
26. Branch DW, Wheeler BC, Brewer GJ, Leckband DE. Long-Term Stability of Grafted Polyethylene Glycol Surfaces for use with Microstamped Substrates in Neuronal Cell Culture. Biomaterials, 2001;22:1035–1047.
27. Alcantar NA, Aydil ES, Israelachvili JN. Polyethylene Glycol-Coated Biocompatible Surfaces. J Biomed Mat Res 2000;51:343–351.
28. Baltes H, Paul O, Brand O. Micromachined Thermally Based CMOS Microsensors. Proc IEEE 1998;86:1660–1678.
29. Stotts LJ. Introduction to Implantable Biomedical IC Design. IEEE Circuits Devices Mag 1989;5:12–18.
30. Stouraitis T, Paliouras V. Considering the Alternatives in Low-Power Design. IEEE Circuits Devices Mag 2001;17:22–29.
31. Tsividis Y, Krishnapura N, Palakas Y, Toth L. Internally Varying Analog Circuits Minimize Power Dissipation. IEEE Circuits Devices Mag 2003;19:63–72.
32. Benini L, De Micheli G, Macii E. Designing Low-power Circuits: Practical Recipes. IEEE Circuits Systems Mag 2001;1:6–25.
33. Rajput SS, Jamuar SS. Low Voltage Analog Circuit Design Techniques. IEEE Circuits Systems Mag 2002;2:24–42.
34. Mohseni P, et al. An Ultra-Light Biotelemetry Backpack for Recording EMG Signals in Moths. IEEE Trans Biomed Eng. June 2001;48:734–737.
35. Proakis JG, Salehi M. Communication System Engineering. Pearson Education; 2001.
36. Lee TH. Design of CMOS Radiofrequency Integrated Circuits. Cambridge: Cambridge University Press, 1998.
37. Razavi B. Challenges in Portable RF Transceiver Design. IEEE Circuits Devices Mag 1996;12:12–25.
38. Larson LE. Integrated Circuit Technology Options for RFICs-Present Status and Future Directions. IEEE J Solid-State Circuits 1998;33:387–399.
39. Crow BP, Wudiaja I, Kim LG, Saki PT. IEEE 802.11 Wireless Local Area Networks. IEEE Commun Mag 1997;35:116–126.
40. Chatschik B. An Overview of the Bluetooth Wireless technology. IEEE Commun Mag 2001;39:86–94.
41. Saltzstein WE. Bluetooth and Beyond: Wireless Options for Medical Devices. Med Device Diagnostic Ind June 2004.
42. Troyk P. Injectable Electronic Identification, Monitoring, and Stimulation Systems. Ann Rev Biomed Eng 1999;1:177–209.
43. Finkenzeller K. RFID Handbook, New York: John Wiley & Sons, Inc; 2003.
44. Kraus JD. Antenna. New York: McGraw-Hill; 2001.
45. Woodward B, Istepanian RSH.Acoustic Biotelemetry of Data from Divers, Proc 15th Annu Int IEEE Eng Med Biol Soc Conf Paris 1992;1000–1001.
46. Kawahito S, et al. A CMOS Integrated Circuit for Multichannel Multiple-Subject Biotelemetry using Bidirectional Optical Transmissions. IEEE Trans Biomed Eng 1994;41: 400–406.

47. Linden D, Reddy T. Handbook of Batteries. New York: McGraw-Hill; 2001.
48. Foster KR, Schwan HP. Handbook of Biological Effects of Electromagnetic Fields, In: Polk C, Postow E, editor. Boca Raton (FL): CRC Press; 1996.
49. Ko WH, Liang SP, Fung CDF. Design of Radio-Frequency Powered Coils for Implant Instruments. Med Biol Eng Computing 1977;15:634–640.
50. Ashby KB, et al. High Q Inductors for Wireless Applications in a Complementary Silicon Bipolar Process. IEEE J Solid-State Circuits 1996;31:4–9.
51. Sokal NO, Sokal AD. Class E-A New Class of High-Efficiency Tuned Single-Ended Switching Power Amplifiers. IEEE J. Solid-State Circuits 1975;10:168–176.
52. Ziaie B, Rose SC, Nardin MD, Najafi K. A Self-Oscillating Detuning-Insensitive Class-E Transmitter for Implantable Microsystems. IEEE Trans Biomed Eng 2001;48:397–400.
53. Mehta V, Cooper JS. Review and Analysis of PEM Fuel Cell Design and Manufacturing. J Power Sources 2003;114:32–53.
54. Singh D, et al. Challenges in Making of Thin Films for $Li_xM_nyO_4$ Rechargeable Lithium Batteries for MEMS. J Power Sources 2001;97–98:826–831.
55. Lal A, Blanchard J. Daintiest Dynamos: Nuclear Microbatteries. IEEE Spectrum 2004;42:36–41.
56. Starner T. Human Powered Wearable Computing. IBM J Systems 1996;35:618–629.
57. Ratner BD, Schoen FJ, Hoffman AS, Lemons JE. Biomaterials Science: An Introduction to Materials in Medicine. New York: Elsevier Books; 1997.
58. Loeb GE, Bak MJ, Salcman M, Schmidt EM. Parylene C as a Chronically Stable reproducible Microelectrode material. IEEE Trans Biomed Eng 1977;24:121–128.
59. Nichols MF. The Challenges for Hermetic Encapsulation of Implanted Devices. Critical Rev Biomed Eng 1994;22:39–67.
60. Schmidt MA. Wafer-to-Wafer Bonding for Microstructure Formation. Proc IEEE 1998;86:1575–1585.
61. Mokwa W, Schenakenberg U. Micro-Transponder Systems for Medical Applications. IEEE Trans Instr Meas 2001;50:1551–1555.
62. Stangel K, et al., A Programmable Intraocular CMOS Pressure Sensor System Implant. IEEE J Solid-State Circuits 2001;36:1094–1100.
63. Iddan G, Meron G, Glukhovsky A, Swain P. Wireless Capsule Endoscopy. Nature (London) 2000;405:417.
64. http://www.givenimaging.com.
65. Santini JT, Cima MJ, Langer R. A Controlled-Release Microchip. Nature (London) 1999;397:335–338.
66. Santini JT, et al. Microchips as Controlled Drug-Delivery Devices. Angew Chem 2000;39:2396–2407.
67. Available at http://www.mchips.com.
68. Available at http://www.chiprx.com.
69. Lei M, et al. A Hydrogel-Based Wireless Chemical Sensor. Proc IEEE MEMS 2004;391–394.
70. Von Arx JA, Najafi K. A Wireless Single-Chip Telemetry-Powered Neural Stimulation System. IEEE Solid-State Circuits Conf 1999;15–17.
71. Liu W, et al. Retinal Prosthesis to Aid the Visually Impaired. IEEE Systems, Man, and Cybernetics, Conf 1999;364–369.
72. Humayun MS, et al. Towards a Completely Implantable, Light-Sensitive Intraocular Retinal Prosthesis. Proc 23rd Ann IEEE EMBS Conf 2001;3422–3425.

See also BIOFEEDBACK; BLADDER DYSFUNCTION, NEUROSTIMULATION OF; MONITORING, INTRACRANIAL PRESSURE; NEONATAL MONITORING; PACEMAKERS.

BIRTH CONTROL. See CONTRACEPTIVE DEVICES.

BLEEDING, GASTROINTESTINAL. See GASTROINTESTINAL HEMORRHAGE.

BLADDER DYSFUNCTION, NEUROSTIMULATION OF

MAGDY HASSOUNA
Toronto Western Hospital
NADER ELMAYERGI
MAZEN ABDELHADY
McMaster University

INTRODUCTION

The discovery of electricity introduced enormous changes to human society: Electricity not only improved daily life, but also opened up new opportunities in scientific research. The effects of electrical stimulation on muscular and nervous tissue have been known for several centuries, but the underlying electrophysiological theory to explain these effects was first derived after the development of classical electrodynamics and the development of nerve cell models (1).

Luigi Galvani first suggested that electricity could produce muscular contraction in his animal experiments (2). He found that a device constructed from dissimilar metals, when applied to the nerve or muscle of a frog's leg, would induce muscular contraction. His work formed the foundation for later discoveries of transmembrane potential and electrically mediated nerve impulses. Alessandro Volta, the inventor of the electrical battery (or voltaic pile) (3), was later able to induce a muscle contraction by producing a potential with his battery and conducting it to a muscle strip. The use of Volta's battery for stimulating nerves or muscles became known as galvanic stimulation.

Another basis for modern neural stimulators was the discovery of the connection between electricity and magnetism, demonstrated by Oersted in 1820; he described the effect of current passing through a wire on a magnetized needle. One year later, Faraday showed the converse—that a magnet could exert a force on a current-carrying wire. He continued to investigate magnetic induction by inducing current in a metal wire rotating in a magnetic field. This device was a forerunner of the electric motor and made it possible to build the magneto-electric and the induction coil stimulator. The latter, the first electric generator, was called the Faraday stimulator. Faradic stimulation could produce sustained titanic contractions of muscles, instead of a single muscle twitch as galvanic stimulation had done.

Duchenne used an induction coil stimulator to study the anatomy, physiology, and pathology of human muscles. Finally, he was able to study the functional anatomy of individual muscles (4,5). This work is still valid for the investigation of functional neuromuscular stimulation.

Another basis for modern stimulator devices lay in the work of Chaffee and Light (6). They examined the problem

of stimulating neural structures deep in the body, while avoiding the risk of infection from percutaneous leads: They implanted a secondary coil underneath the skin and placed a primary coil outside the body, using magnetic induction for energy transfer and modulation. Further improvement was achieved by radio frequency (rf) induction (7,8). The Glenn group developed a totally implanted heart pacemaker—one of the first commercially available stimulators. In the ensuing years, stimulators for different organ systems were developed, among them the abovementioned heart pacemaker, a diaphragmatic pacemaker (7,8), and the cochlear implant (9).

BLADDER STIMULATION

Electrical stimulation of the bladder dates back to 1878. The Danish surgeon M.H. Saxtorph treated patients with urinary retention by inserting a special catheter with a metal electrode into the urinary bladder transurethrally and placing a neutral electrode suprapubically (10). Also, Katona et al. (11) described their technique of intraluminal electrotherapy, a method that was initially designed to treat a paralytic gastrointestinal tract, but was later used for neurogenic bladder dysfunction in patients with incomplete central or peripheral nerve lesions (11,12).

Further interest in the electrical control of bladder function began in the 1950s and 1960s. The most pressing question at that time was the appropriate location for stimulation. Several groups attempted to initiate or prevent voiding (in urinary retention and incontinence, respectively) by stimulation of the pelvic floor, the detrusor directly, the spinal cord, or the pelvic and sacral nerves or sacral roots. Even other parts of the body, such as the skin, were stimulated in an attempt to influence bladder function (13).

In 1954, McGuire performed extensive experiments of direct bladder stimulations in dogs (14) with a variety of electrodes, both single and multiple, in a variety of positions. Boyce and associates continued this research (15).

It was realized that with a single pair of electrodes, the maximal response was obtained when the electrodes were placed on both lateral bladder walls so that the points of stimulation encompassed a maximal amount of detrusor muscle. When this was performed in human studies, an induction coil for direct bladder stimulation was implanted in three paraplegic men with complete paralysis of the detrusor muscle. The secondary coil was implanted in the subcutaneous tissue of the lower abdominal wall. Of the three, only one was a success, with the other a failure and the third only partially successful (15).

In 1963, Bradley and associates published their experience with an implantable stimulator (16). They were able to achieve complete bladder evacuation in the chronic dog model over 14 months. However, when the stimulator was implanted in seven patients, detrusor contraction was produced, but bladder evacuation resulted in only two. Further experiments were performed in the sheep, calf, and monkey in an attempt to resolve species discrepancies. These animals were chosen because, in the sheep and calf, the bladder is approximately the same size as in the human, and this similarity could determine whether more power is needed for a bladder larger than that of the dog. In addition, the pelvis of monkeys and humans is similarly deep; thus, the influence (if any) of pelvic structure could be investigated. The results showed that a larger bladder needs more power and wider contact between the electrodes and that differences in structure do not necessitate different stimulation techniques (13,16).

PELVIC FLOOR STIMULATION

In 1963, Caldwell described his clinical experience with the first implantable pelvic floor stimulator (17). The electrodes were placed into the sphincter, with the secondary coil placed subcutaneously near the iliac spine. Though this device was primarily designed for the treatment of fecal incontinence; Caldwell also treated urinary incontinence successfully.

Another approach to pelvic floor stimulation for females is intravaginal electrical stimulation, reported initially by Magnus Fall's group (1977) (18). They published numerous studies dealing with this subject in the ensuing years and found that intravaginal electrical stimulation also induces bladder inhibition in patients with detrusor instability. Lindstram, a member of the same group, demonstrated that bladder inhibition is accomplished by reflexogenic activation of sympathetic hypogastric inhibitory neurons and by central inhibition of pelvic parasympathetic excitatory neurons to the bladder (13,19). The afferent pathways for these effects could be shown to originate from the pudendal nerves.

POSTERIOR TIBIAL OR COMMON PERONEAL

Another interesting application of electrical stimulation for inhibition of detrusor activity is the transcutaneous stimulation of the posterior tibial or common peroneal nerve. This technique, drawn from traditional Chinese medicine, is based on the acupuncture points for inhibition of bladder activity and was reported by McGuire et al. in 1983 (20).

A percutaneous tibial nerve stimulation (PTNS) (Urgent PC, CystoMedix, Anoka, MN) was approved by the Food and Drug Administration in 2000. A needle is inserted $\sim$5 cm cephalad from the medial malleolus and just posterior to the margin of the tibia. Stimulation is done using a self-adhesive surface stimulation electrode without an implanted needle electrode (21). Current data describe results after an initial treatment period of 10–12 weeks. If patients get a good response, they are offered tapered chronic treatment. As in sacral root neuromodulation, PTNS seems less effective for treating chronic pelvic pain (22).

More substantial data, in particular on objective parameters and long-term follow up, are needed, as are studies looking into the underlying neurophysiological mechanisms of this treatment modality. Although minimally invasive, easily applicable, and well tolerated, the main disadvantage of PTNS seems to be the necessity of chronic treatment. The development of an implantable subcutaneous stimulation device might ameliorate this problem (23). It has never found widespread acceptance.

PELVIC NERVE STIMULATION

Pelvic nerves do not tolerate chronic stimulation and the pudendal nerves are activated, increasing outflow resistance. Also, in humans the fibers of the parasympathetic nervous system innervating the bladder split early in the pelvis, forming a broad plexus unsuitable for electrode application (24).

DETRUSOR STIMULATION

Direct detrusor stimulation offers high specificity to the target organ (25), but its disadvantages are electrode displacement and malfunction due to bladder movement during voiding, and fibrosis (even erosion) of the bladder wall. In 1967, Hald et al. (26) reported their experience of direct detrusor stimulation with a radio-linked stimulator in four patients, three with upper motor-neuron lesions and one with a lower motor-neuron lesion. The receiver was placed in a paraumbilical subcutaneous pocket. Two wires from the receiver were passed subcutaneously to the ventral bladder wall, where they were implanted. A small portable external transmitter generated the necessary energy. The procedure worked in three patients; in one it failed because of technical problems (13).

SPINAL CORD STIMULATION

The first attempt to achieve micturition via spinal cord stimulation was through the exploration of the possibility of direct electrical activation of the micturition center in the sacral segments of the conus medullaris. This was conducted by Nashold, Friedman, and associates, and had reported that the region for optimal stimulation was S1–S3.

Effectiveness was determined not only by location, but also by frequency. In two subsequent experiments, the same group compared the stimulation of the dorsal surface of the spinal cord at LS, S1, and S2 with depth stimulation (2–3 mm) at S1 and S2 in acute and chronic settings (27). It was only through the latter, the depth stimulation, that voiding was produced: High bladder pressures were achieved by surface stimulation, but external sphincter relaxation did not occur, and was noted only after direct application of the stimulus to the micturition center in the spinal cord. Stimulation between L5 and S1 produced pressure without voiding, even with depth stimulation (13).

Jonas et al. continued the investigation of direct spinal cord stimulation to achieve voiding (28–30). They compared 12 different types of electrodes: three surface (bipolar surface electrode, dorsal column electrode, and wrap-around electrode) and nine depth electrodes. These differed in many parameters (e.g., bipolar–tripolar, horizontal–vertical–transverse). Regardless of the type of electrode, the detrusor response to stimulation was similar. Interestingly, the wrap-around surface electrode with the most extended current spread provoked the same results as the coaxial depth electrode with the least current spread, prompting those authors to theorize that current does not cross the midline of the spinal cord. Unfortunately, no real voiding was achieved. It was found that the stimulation of the spinal cord motor centers stimulates the urethral smooth and striated sphincteric elements simultaneously: The expected detrusor contraction resulted, but sphincteric contraction was associated. The sphincteric resistance was too high to allow voiding: It allowed only minimal voiding at the end of the stimulation, so-called poststimulus voiding (13). These results contrasted with the earlier work of Nashold and Friedman (27,31).

Thurhoff et al. (32) determined the existence of two nuclei, a parasympathetic and a pudendal nucleus. The parasympathetic nucleus could be shown within the pudendal nucleus; thus, at the level of the spinal cord, stimulation of the bladder separate to that of the sphincter is difficult.

SACRAL ROOT STIMULATION

Based on the hypothesis that different roots would carry different neuronal axons to different locations. The culmination of these studies led to the feasibility of sacral rootlet stimulation.

It appears that sacral nerve-root stimulation is the most attractive method since the space within the spinal column facilitates mechanically stable electrode positioning and the application of electrodes is relatively simple due to the long intraspinal course of the sacral roots.

The University of California, San Francisco (UCSF) group performed numerous experiments on a canine model (33), as the anatomy of bladder innervation of the dog is similar to that of the human. After laminectomy, the spinal roots were explored and stimulated, either intradurally or extradurally, but within the spinal canal, in the following modes:

1. Unilateral stimulation of the intact sacral root at various levels.
2. Simultaneous bilateral stimulation of the intact sacral root at various levels.
3. Stimulation of the intact ventral and dorsal root separately.
4. Stimulation of the proximal and distal ends of the divided sacral root.
5. Stimulation of the proximal and distal ends of the divided dorsal and ventral roots (13).

From these studies, it became clear that stimulating the intact root is least effective and stimulating the ventral component is most effective and that no difference exists between right- and left-root stimulation (33).

However, this stimulation also causes some sphincteric contraction, owing to the presence of both autonomic and somatic fibers in the ventral root, and the studies were continued with the addition of neurotomy to eliminate the afferent fibers. These experiments showed that, to achieve maximally specific detrusor stimulation, the dorsal component must be separated from the ventral component and the somatic fibers of the root must be isolated and selectively cut (34).

The experiments also demonstrated that stimulation with low frequency and low voltage can maintain adequate sphincteric activity, but that stimulation with high frequency and low voltage will fatigue the external sphincter and block its activity. When high frequency/low voltage stimulation is followed by high voltage stimulation, bladder contraction will be induced and voiding achieved.

The finding that detrusor contraction can be activated separately from sphincteric activity and that adequate sphincteric contraction can be sustained without exciting a detrusor reaction made it seem possible that a true bladder pacemaker could be achieved. In addition, in histological and electron microscopic examination of the stimulated sacral roots, no damage was found when they were compared with the contralateral nonstimulated roots. Neither the operation nor the chronic stimulation damaged the ventral root, and the responses remained reliable and stable (13).

Tanagho's group later performed detailed anatomical studies on human cadavers. The aim was to establish the exact anatomical distribution of the entire sacral plexus, following it from the sacral roots in the spinal cord through the sacral foramen inside the pelvic cavity. Emphasis was placed on the autonomic pelvic plexus as well as the somatic fibers. With this anatomical knowledge, the stimulation of human sacral roots in neurogenic bladder dysfunction was developed and made clinically applicable as a long-term treatment (35). Direct electrical stimulation was performed through a permanently implanted electrode, placed mostly in contact with S3 nerve roots in the sacral foramen, after deafferentation.

The stimulation of sacral rootlet bundles isolated from the rest of the sacral root gave the same increase of bladder pressure when stimulated close to the exit from the dura, in the mid-segment, or close to the origin in the spinal cord. This could make the stimulation more selective, eliminating detrusor-sphincter dyssynergia.

In additional work, taking advantage of the knowledge that high frequency current can block large somatic fibers, electrical blockade of undesired responses was tested to replace selective somatic neurotomies. High frequency sinusoidal stimulation was effective in blocking external sphincter activity. However, the sinusoidal waveform is not efficient. Alternate-phase, rectangular wave is more efficient and induces the same blockade: alternating pulses of high frequency and low amplitude followed by pulses of low frequency and high amplitude were effective in inducing low pressure voiding without the need for somatic neurotomies. This approach has not yet been tried clinically, but it might prove to be the answer to the problem of detrusor-sphincter dyssynergia in electrically stimulated voiding (13).

The three main devices used for sacral neuromodulation is the Medtronic InterStim, the Finetech–Brindley (VOCARE) bladder system, and the rf BION systems. Each is explained in detail below.

MEDTRONIC INTERSTIM

Indications for use: urge incontinence, retention and urgency frequency, male and female dysfunctional voiding syndromes and postprostatectomy incontinence. There are also benefits beyond voiding disorders, including re-establishment of pelvic floor awareness, resolution of pelvic floor muscle tension and pain, reduction in bladder pain (interstitial cystitis) and normalization of bowel function.

The basic concept behind the implantable pulse generator (IPG) that provides stimulation to the sacral nerve is not far removed from the concepts behind cardiac pacing. A long-lived battery encased in biocompatible material is programmed to deliver pulses of electricity to a specific region of the body through an electrode at the end of an encapsulated wire.

Medtronic is the manufacturer of the InterStim neurostimulator. Earl Bakken, the founder of the company, first created a wearable, battery-operated pacemaker at the request of Dr. C. Walton Lillehei, a pioneer in open-heart surgery at the University of Minnesota Medical School Hospital, who was treating young patients for heart block.

The Itrel I, the first-generation neurostimulator, was introduced in 1983. Current versions are used for the treatment of incontinence, pain, and movement disorders.

System Overview

There are two established methods for sacral root neuromodulation using the Medtronic InterStim system.

1. An initial test phase, then the more permanent hardware is implanted.
2. An alternative method uses a staged testing–implant procedure, where a chronic lead is implanted and connected to a percutaneous extension and test stimulator.

Testing Phase (See Fig. 1). The testing hardware consists of a needle, test lead, test stimulator, interconnect cabling and a ground pad (Fig. 1).

- Needle (see Figs. 2 and 3).

A 20-gauge foramen needle with a bevelled tip is used to gain access to the sacral nerve for placing the test stimulation lead. The stainless steel needle is depth-marked along its length and electrically insulated along its center length. The portion near the hub is exposed to allow connection to

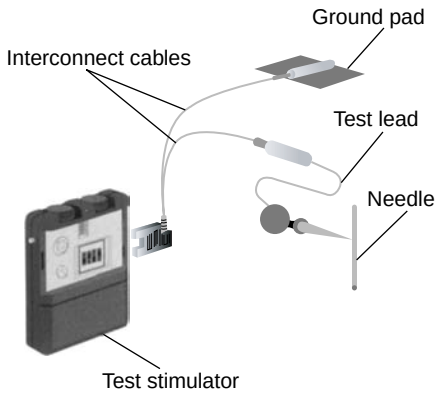


Figure 1. Test stimulation system.

Figure 2. Model 041828 (20 gauge) 3.5 in. (88.9 mm) foramen needles.

Figure 3. Model 041829 (20 gauge) 5 in. (127 mm) foramen needles.

the test stimulator. By stimulating through the uninsulated tip of the needle, the physician can determine the correct SNS site for the test stimulation lead.

- Test lead (see Fig. 4).

The initial test lead is a peripheral nerve evaluation (PNE) test lead with a coiled, seven-stranded stainless steel wire coated with fluoropolymer. Its electrode is extended to 10 mm (0.4 in.) to increase the length of coverage and reduce the effects of minor migration. Depth indicators help to align the lead electrode with the needle tip. The lead contains its own stylet, which is removed once the correct position has been found, leaving the lead flexible and stretchable, to mitigate migration.

- Test stimulator (see Fig. 5).

The most current version of test stimulators is the model 3625. The model 3625 test stimulator can be used both for patient screening, where the patient is sent home with the device, and for intraoperative usage in determining lead placement thresholds. It provides output characteristics that are similar to those of the implantable neurostimulator and can be operated in either monopolar or bipolar modes. It is battery operated by a regular, 9 V battery. The physician sets the maximum and minimum amplitude settings, allowing the patient to control the amplitude (within those maximum and minimum settings) to whatever level is comfortable.

The safety features of the stimulator include; an automatic output shut-off occurs when the amplitude is turned up too rapidly (as when the control is inadvertently bumped), a loose device battery will cause output shut-off also to prevent intermittent stimulation and shock to the patient, and sensors, which detect when electrocautery is being used, shut the output off. Turning the test stimulator off for a minimum of 3 s can reset the protection circuitry.

- Interconnect cables (see Fig. 6).

Single-use electrical cables are used to hook the test stimulation lead to the model 3625 test stimulator during the test stimulation procedure in the physician's office and when the patient goes home for the evaluation period.

Figure 4. Model 3057 test stimulation lead.

Figure 5. Model 3625 sacral nerve test stimulator.

The patient cable is used to deliver acute stimulation during the test procedure. The insulated tin-plated copper cable has a 2 mm socket at one end and a spring-activated minihook at the other end. The minihook makes a sterile connection to the foramen needle, test stimulation lead, or implant lead. The socket end is connected to the test stimulator by a long screener cable, the latter being a two-wire cable with a single connector to the model 3625 test stimulator at one end; one of the wires is connected to the patient cable and the other to the ground pad. After the test stimulation, the patient cable is removed and a short screener cable is substituted for at-home use. This cable is connected to the ground pad and directly to the test lead. It is designed to withstand the rigours of home use and can be disconnected, to facilitate changing clothes (13).

- Ground pad.

The ground pad provides the positive polarity in the electrical circuit during the test stimulation and the at-home trial. It is made of silicone rubber and is adhered to the patient's skin. As described above, for the at-home trial a short screener cable is substituted for the long screener cable and connected directly to the lead.

Surgical Technique Used for Acute Testing Phase:. The aims of percutaneous neurostimulation testing (PNE) are to check the neural and functional integrity of the sacral nerves, to determine whether neurostimulation is beneficial for each particular patient, and to clarify which sacral spinal nerves must be stimulated to achieve the optimum therapeutic effect in each individual case.

Local anesthetic is injected into the subcutaneous fatty tissue and the muscles, but not into the sacral foramen itself. The S3 foramen is localized on one side with a 20-gauge foramen needle. By stimulating through the uninsulated tip of the needle, the physician can find the correct sacral nerve stimulation site for placement of the test stimulation lead. Once the location of the S3 foramen is established, tracing of the other foramina is done. The

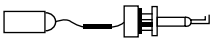


Figure 6. Model 041831 patient cable.

portion near the hub is exposed to allow connection to the test stimulator.

Keeping the needle at a 60° angle to the skin surface with a rostrocaudal and slightly lateral pointing tip of the needle will ensure that the needle is inserted into the targeted foramen. The puncture should progress parallel to the course of the sacral nerve, which normally enters at the upper medial margin of the foramen. This method achieves optimal positioning of the needle for stimulation and avoids injuring the spinal nerve. The insulated needle (cathode) is then connected to an external, portable pulse generator (Medtronic model 3625 test stimulator) via a connection cable. The pulse generator itself is connected to a neutral electrode (anode) attached to the shoulder.

Because patient sensitivity varies, the voltage used is between 1–6 V, which starts at 1 and is increased in 20 Hz increments. Stimulation of the S3 evokes the "bellows" effect (contraction of the levator ani and the sphincter urethrea). Also, there is plantar flexion of the foot on the ipsilateral side. If plantar flexion of the entire foot is observed, the gastrocnemius muscle should be palpated, because a strong contraction usually indicates stimulation of S2 fibers and should be avoided.

Stimulation of S3 generally produces the most beneficial effect. Furthermore, most patients will not tolerate the permanent external rotation of the leg caused by stimulation of S2. Occasionally, stimulating S4 also causes clinical improvement. Stimulation of S4 provokes a strong contraction of the levator ani muscle, accompanied by a dragging sensation in the rectal region. If stimulating one side produces an inadequate response, the contralateral side should be tested; the aim is to obtain a typical painless stimulatory response.

Once the optimal stimulation site has been identified, the obturator is removed from the foramen needle, and a temporary wire test lead (Medtronic model 3057 test lead) is inserted through the lumen of the needle. Once the test lead has been inserted into the needle, the latter must not be advanced any further in order to avoid severing the lead. The needle is then carefully removed from the sacral foramen, leaving the test lead in place. The stimulation is then repeated to check the correct position of the test electrode. To mitigate migration the lead contains its own stylet, which is removed once the correct position has been found, leaving the lead flexible and stretchable.

A repetition of the test stimulation, confirming the correct position of the test lead, is therefore mandatory at this stage; otherwise the test lead cannot be reinserted.

After correct positioning, the test lead is coiled on the skin and fixed with adhesive transparent film. Finally, the correct position of the wire is radiologically confirmed and the portable external impulse generator is connected.

Percutaneous Extension Hardware (see Fig. 7). If acute testing is inconclusive, or when there is a need for positive fixation of the test lead, percutaneous extension hardware is the best method used. Also called the staged implant, it is an alternative method for patient screening.

The chronic lead is implanted in the normal manner and is connected to a percutaneous extension (model 3550-05). The extension is designed to provide a connection between the chronic lead and the external test stimulator. Positive contact is made using four set screws; the connection is sealed with a silicone boot that covers the set screws. The percutaneous extension, which is intended for temporary use, features four insulated wires, wound together and sized for a small incision, so that they can be brought through the skin. The percutaneous extension is then connected to the screener cable, as described above (13).

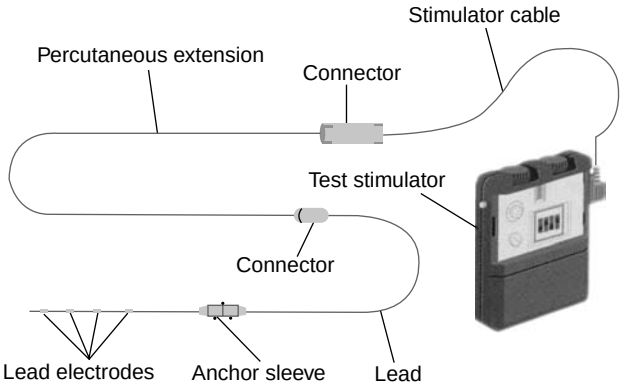


Figure 7. Percutaneous extension system.

Chronic System. The chronic system consists of an implantable neurostimulator, a lead, an extension, a physician programmer and a patient programmer.

- Neurostimulator (see Fig. 8).

The implantable neurostimulator (Medtronic model 3023) weighs ~42 g and has a volume of 22 cm^3. It comprises ~70% battery and 30% electronics. The physician has unlimited access to programmable parameters such as amplitude, frequency, and pulse width. Each parameter can be changed by means of an external, physician programmer that establishes a rf link with the implanted device. A patient programmer provides limited access to allow the patient to turn the neurostimulator on and off, or to change amplitude within a range established by the physician (via the physician programmer) (13).

The external titanium container of the neurostimulator may be used in either a monopolar configuration (lead negative, can positive) or a bipolar configuration, which will result in marginally better longevity. The life of the neurostimulators is usually ~7–10 years. Factors that affect this are the mode, programming of the amplitude, pulse width and frequency, and the use of more than one active electrode.

- Implantable lead system (see Fig. 9–14).

Figure 8. Model 3023 implantable neurostimulator.

Figure 9. Model 3080 lead.

Figure 10. Model 3092 lead.

Figure 11. Model 3093 lead.

Figure 12. Model 3886 lead.

Figure 13. Model 3889 lead.

Figure 14. Model 3966 lead.

The lead is a quadripolar design, with four separate electrodes that can be individually programmed to plus, minus, or off. This allows the physician to optimize the electrode configuration for each patient and to change programming, without additional surgery, at a later date, to adapt to minor lead migration or changing disease states. The electrode sizes, spacing, and configurations have been designed specifically for SNS.

The lead is supplied with multiple stylets and anchors, to accommodate physician preferences. A stylet (straight or bent) is inserted into the lumen of the lead to provide extra stiffness during implant. Two different degrees of stiffness provide the physician with options to tailor the handling and steering properties of the lead, as preferred. The stylet must be removed before connection with the mating component.

The physician also has a choice of anchors, which allow fixation of the lead to stable tissue to prevent dislodging of the lead after implantation. Three anchor configurations are available: a silicone rubber anchor fixed in place on the lead has wings, holes and grooves to facilitate suturing; a second type, also made of silicone, slides into place anywhere along the lead body, and must be sutured to the lead to hold it in place; a new plastic anchor is also available, which can be locked in place anywhere along the lead body without a suture to the lead.

- Quadripolar extension (see Fig. 15).

Figure 15. Series 3095 extension.

Figure 16. Physician programmer.

The quadripolar extension, which is available in varying lengths to facilitate flexibility in IPG placement, is designed to provide a sealed connection to the lead. This extension provides the interface with the neurostimulator. Positive contact is made with four set screws, and the connection is sealed with a silicone boot covering the screws.

- Physician programmer (Fig. 16).

The console programmer (Medtronic model 8840 N'Vision) is a microprocessor-based system that the physician uses to program the implanted neurostimulator noninvasively. The programmer uses an application-specific memory module, installed by means of a plug-in software module.

- Patient programmer (Fig. 17).

The patient programmer also communicates with the implanted neurostimulator by an rf link. The patient can adjust stimulation parameters within the range set by the physician. This range is intended to allow the patient to turn the device on or off, and to change amplitude for comfort (as during postural changes), without returning to the physician's office.

Surgical Technique Used for Chronic Implantable System. The sacral foramen electrode and impulse generator are implanted under general anesthesia. Long-acting muscle relaxants must not be used, as these would impair the intraoperative electrostimulation.

The patient is placed in the prone position with a 45° flexion of the hip and knee joints. An 8 cm long midline

Figure 17. Patient programmer.

incision is made above the sacrum, reaching one-third caudal and two-thirds cranial from the S3 foramen. After transection of the subcutaneous fat, the muscle fascia (thoracolumbar fascia) is incised approximately 1 cm lateral of the midline in a longitudinal direction.

Usually, the Gluteus maximus has to be incised over a length of 1–2 cm for good exposure of the S3 foramen and a little further caudal if implantation of the S4 foramen is intended. The paraspinal muscles are then divided longitudinally and the dorsal aspect of the sacrum is exposed.

Intraoperative test stimulation, using the same equipment as for the acute testing phase, will confirm the precise location of the foramen selected. The foramen needle is left in place to avoid relocalisation of the foramen while preparing the permanent electrode for implantation. Proximal to the four contact points of the permanent electrode, a silicon rubber cuff is glued to the electrode body. The cuff is fitted with three eyelets to accommodate nonabsorbable atraumatic needle-armed sutures.

After removal of the foramen needle, the permanent electrode (Medtronic quadripolar lead, model 3080) is gently inserted into the foramen. Renewed test stimulation will determine the most effective contact point between the electrode and spinal nerve; the most distal contact point is termed "0", with the subsequent three being numbered 1–3 sequentially. An identical motor response at all four contact points is ideal. If only one contact gives a satisfactory response, the electrode should be repositioned at a different angle to the foramen and the test stimulation repeated. The preattached sutures are then used to secure the electrode to the ligaments overlying the periosteum of the sacral bone. Test stimulation should be repeated at this stage to confirm an appropriate position of the electrode after fixation.

A small skin incision is now made in the flank between the iliac crest and the 12th rib on the side where the electrode has been placed. A subcutaneous tunnel is formed between the two wounds, starting from the flank incision and running toward the sacral incision.

The obturator of the tunneling device is removed and the silicone sheath, which is open at both ends, left in place. The free end of the electrode is guided through the sheath to the flank incision, after the stylet has been removed from the electrode.

The silicone sheath is now removed from the flank incision, the proximal end of the electrode is marked with a suture, and the electrode is buried in a subcutaneous pocket that has been created at the site of the flank incision. The flank incision is temporarily closed, leaving the marking suture exposed between the skin sutures. The sacral incision is then closed in layers and covered with a sterile dressing.

The patient is now positioned on the contralateral flank. The flank and abdomen on the side chosen previously for placement of the Medtronic InterStim model 3023 implantable pulse generator are disinfected and the surgical field is draped with a sterile cover. The flank incision is now reopened, and a subcutaneous tunnel is again created between the flank incision and the subcutaneous pocket in the lower abdomen through which a connecting extension cable (Medtronic quadripolar extension, model 3095) between electrode and impulse generator is guided.

Once the electrode has been connected to the extension cable in the area of the flank incision, the contact point is sealed with a silicone cover, fixed with two sutures and placed subcutaneously. The flank incision is closed in layers and covered with a sterile dressing.

Finally, the other end of the connecting cable is attached to the impulse generator. The generator is attached to the rectus fascia using two nonabsorbable sutures. The abdominal incision is closed in two layers and covered with sterile dressings.

On the first postoperative day, anterior–posterior and lateral radiographs of the implant are obtained to verify that all components are correctly positioned and will act as a control for comparison in case of subsequent problems.

Modifications of the surgical procedure include placement of the pulse generator in the gluteal area thus avoiding repositioning of the patient during the procedure and implantation of bilateral electrodes, which should be powered by a two-channel pulse generator (Medtronic Synergy, model 7427) for adequate synchronous independent stimulation of each side. The implant remains deactivated at least until the day following surgery and will be activated by a telemetric programming unit (Medtronic Console Programmer, model 7432) allowing programming of all features of the implant by the physician during the initial activation and follow-up stages.

NEW MEDTRONIC TINED LEAD PERCUTANEOUS IMPLANT (SEE FIGS. 18 AND 19)

Tined leads offer sacral nerve stimulation through a minimally invasive implant procedure. The use of local anesthesia allows for patient sensory response during the implant procedure. This response helps ensure optimal lead placement and may result in better patient outcomes. With previous lead designs, many physicians used general anesthesia, which did not allow for patient sensory response.

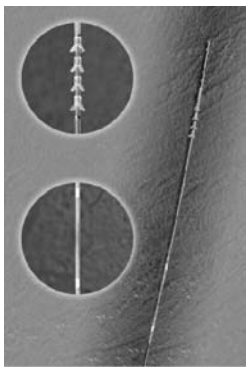


Figure 18. Tined lead percutaneous implant.

Among the advantages of the minimally invasive implant procedure are the radiopaque markers to identify where tines are deployed, helping physicians in identifying the exact lead location relative to the sacrum and nerves. Tactile markers indicate lead deployment, and a white marker bands on the lead and tactile markers aid in proper lead placement and to notify the physician when the tines are ready to be deployed.

Percutaneous lead placement allows use of local anesthesia. This reduces the risks of general anesthesia and surgical incision and may facilitate faster patient recovery time as a result of less muscle trauma and a minimized surgical incision. Also, it may reduce surgical time as a result of a sutureless anchoring procedure and reduced number of surgical steps.

To date, a positive response to the PNE test has been the only predictive factor for the long-term efficacy of sacral nerve stimulation therapy. Current studies show that up to 40% of patients who experience improvement in symptoms during PNE test stimulation with a temporary lead do not have this improvement carried through after neurostimulator implantation (36). A study by Spinelli et al. looked at patients who underwent tined lead implant without PNE testing, and reported a positive outcome of 80% during the screening phase, which was maintained at an average follow up of 11 months, resulting in a higher success rate than that currently reported in the literature (37).

The development of the new tined lead allows fully percutaneous implantation of the permanent lead and offers the possibility of a longer and more reliable screening period than that possible with the PNE test. The advantage for patient screening are that the permanent tined lead is less prone to migration, hence if the results of screening are positive, the lead is already in the precise place where positive results were obtained, and there is a decrease in false-positive and false-negative results after screening (37). However, use (or lack thereof) of PNE testing in conjunction with the tined lead differs from center to center, depending on fiscal and/or other reasons.

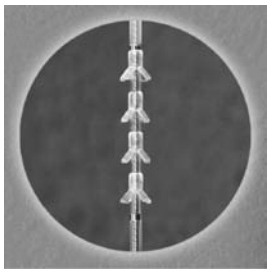


Figure 19. Tined lead percutaneous implant.

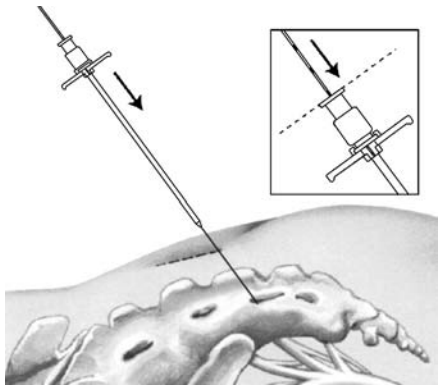


Figure 20. The foramen needle stylet and directional guide.

The tined lead models 3093 and 3889 are designed to work with the current lead introducer model 355018 or 042294.

Surgical Technique for Tined Lead Implant. The foramen needle is inserted and tested for nerve response. The foramen needle stylet is then removed and replaced with the directional guide (see Fig. 20). The foramen needle itself is then removed.

A small incision is made on either side of the directional guide, which is followed by fitting the dilator and the introducer sheath over the directional guide and advanced into the foramen (see Fig. 21). The guide and the dilator are then removed, leaving the introducer sheath in place.

The lead is then inserted into the introducer sheath and advanced until visual marker band C on the lead lines up with the top of the introducer sheath handle. Using fluoroscopy, electrode 0 of the lead is confirmed to be proximal to the radiopaque marker band at the distal tip of the sheath (see Fig. 22).

While holding the lead in place, the introducer sheath is retracted until visual marker band D on the lead lines up with the introducer sheath handle. Using fluoroscopy, radiopaque marker band at the tip of the sheath is

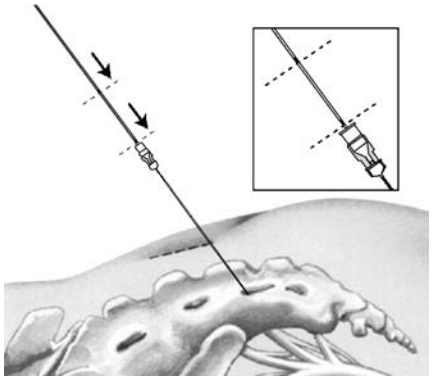


Figure 21. Fitting the dilator and introducer sheath.

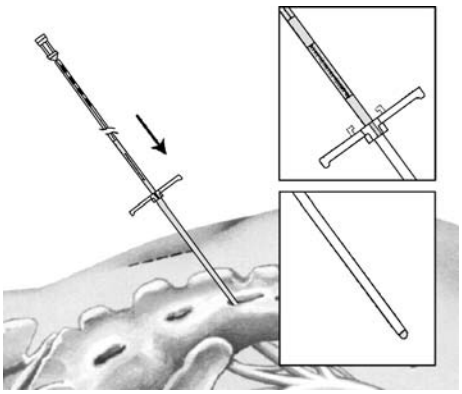


Figure 22. Confirming lead proximal to radiopaque marker band.

confirmed to be proximal to electrode 3 and adjacent to radiopaque marker band A on the lead (see Fig. 23).

Test stimulation of the various electrodes (0, 1, 2, 3) is done and the responses are observed. If necessary, the lead is repositioned within the foramen. When the lead is in the proper position, the lead is held in place and the introducer sheath and lead stylet are carefully withdrawn, thereby deploying the tines and anchoring the lead.

FINETECH–BRINDLEY (VOCARE) BLADDER SYSTEM

Introduction (see Fig. 24)

Indications for use: The VOCARE bladder system is indicated for the treatment of patients who have clinically complete spinal cord lesions with intact parasympathetic innervation of the bladder and are skeletally mature and neurologically stable. However, patients with other neurological disorders, including multiple scleroses, spinal cord tumours, transverse myelitis, cerebral palsy and meningomyelocele, have also benefited from the implant(38). *A secondary use of the device is to aid in bowel evacuation and promote penile erection*.

The sacral anterior root stimulation (SARS) system was developed by Brindley with the support of the Medical Research Council (Welwyn Garden City, Herts, UK), is manufactured by Finetech Medical Ltd. in England, and is marketed as the Vocare system by NeuroControl Corporation (Cleveland, OH) (1).

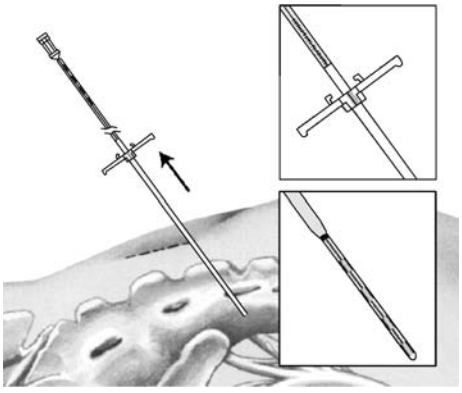


Figure 23. Confirming marker band proximal to electrode 3.

Beginning in 1969, Brindley developed a new device to stimulate sacral roots at the level of the cauda equina. This technique, first tested in baboons, led to the development of a stimulator that was first successfully implanted in a patient in 1978 (39).

Hardware

The Finetech–Brindley bladder controller is composed of external and internal equipment.

1. External (see Fig. 25):

   One analog and three digital versions of the external controller are available in different countries (1). This device has no batteries but is powered and controlled by rf transmission from a portable external controller operated by the user and programmed by the clinician. It consists of a transmitter block connected to the control box via a transmitter lead. The patient holds the transmitter over the implanted receiver to apply stimulation. A new, smaller control box that is more powerful will be available in the coming months (39).
2. Internal (see Fig. 26):

   The internal equipment consists of three main parts: (1) the electrodes, (2) the cables, (3) and the receiver block.

   Two types of electrodes are used, depending on the approach (intra- or extradural).

   For intradural implantation the electrode mounts in which the anterior sacral roots are trapped are called "books" because of their shape.

   The two-channel implant has an upper book with only 1 slot. Trapping of S3 and S4 roots is often sufficient to obtain bladder contractions. In males, S2 roots were trapped in the upper book and S3 and S4 roots, in the lower book.

   The three-channel implant is composed of two electrode books. The upper book contains three parallel slots for S3 and S2 roots and the lower contains one slot for S4 roots. There are three electrodes in each slot (one cathode in the center and two anodes at the two ends) to avoid stimulation of unwanted structures.

   The four-channel implant has two books like those of the three- channel implant, and the four slots allow independent stimulation of four sets of nerve fibers. It is used in patients who retained sacral-segment pain sensitivity.

   The special eight-channel implant allowed the stimulation of four anterior roots and the destruction of any of the four posterior roots, if necessary, after implantation. It is no longer used.

   For extradural implantation the cables end with three helical electrodes (a cathode between two anodes) and are attached to the roots with a strip of Dacron-reinforced silicone rubber. The cables used

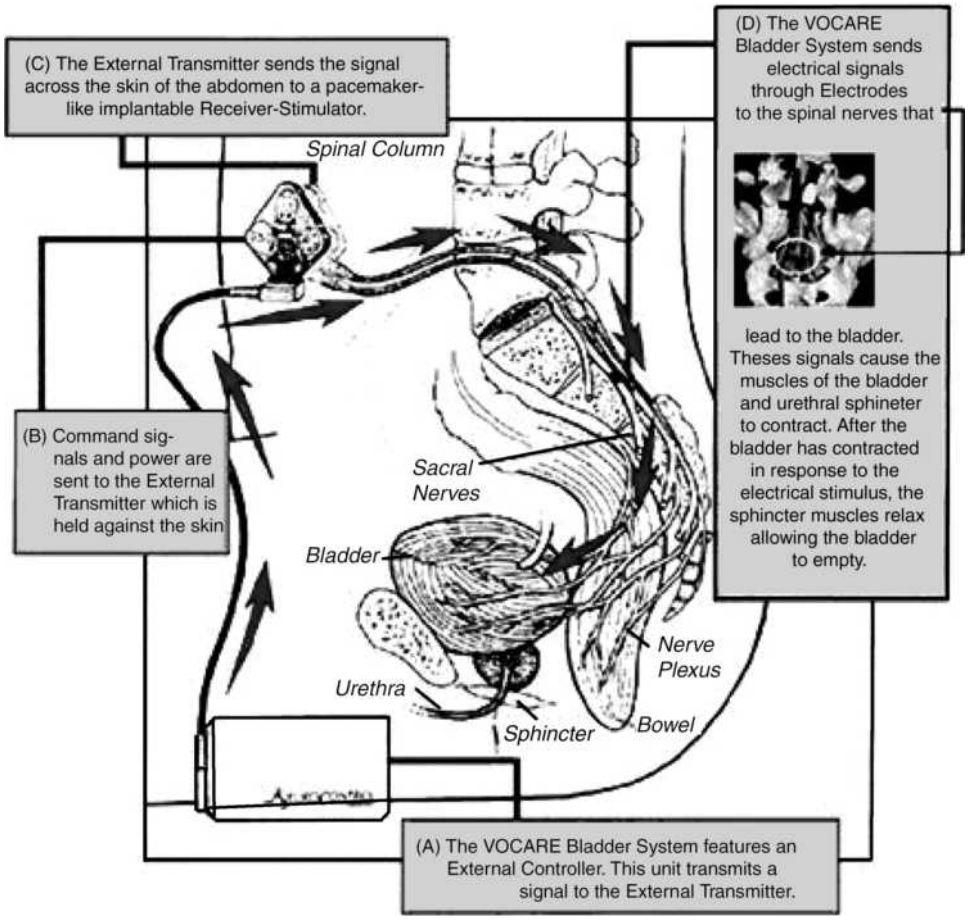


Figure 24. VOCARE bladder system.

Figure 25. External equipment. (a) New control box. (b) Original control box. (c) Transmitter lead. (d) Transmitter block.

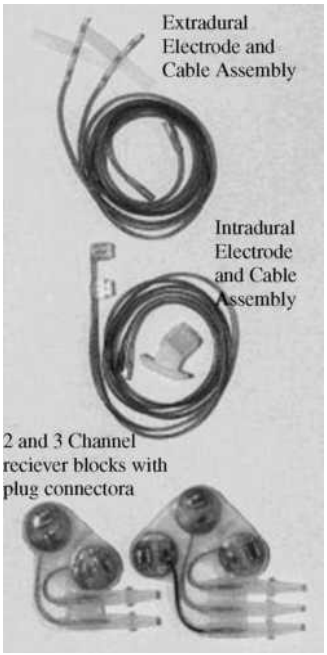


Figure 26. Internal equipment.

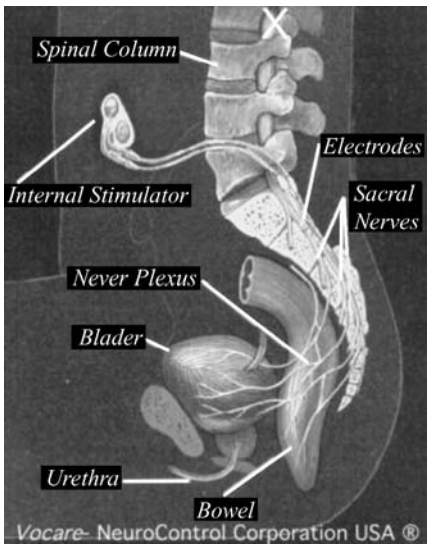


Figure 27. Finetech–Brindley system.

are encapsulated in silicone rubber, and the wires are made of 90% platinum and 10% iridium and connect the electrodes to the radio receiver block. The radio receiver block, which contains two, three, or four radioreceivers imbedded in silicone rubber, is activated by pulse-modulated rf waves (39).

Surgical Technique for Finetech–Brindley System (see Fig. 27):. The surgical technique for ***intrathecal*** implantation developed by Brindley et al. (40) involves laminectomy of the fourth and fifth lumbar vertebrae and the first two pieces of the sacrum, exposing 10–12 cm of dura. The dura and arachnoid are opened at the midline to expose the roots. The roots are identified by their size and situation and by perioperative stimulation during the recording of bladder pressure and observation of skeletal muscle responses with the naked eye.

The S2 anterior roots contract the triceps surae, the glutei, and the biceps femoris. The S3 anterior roots innervate the pelvic floor and the toe flexors. The S4 anterior roots innervate the pelvic floor. The sphincters (anorectal and urethral) are innervated predominantly by S4 and also by S3 and S2. The detrusor response is always obtainable by stimulation of S3 and S4 and sometimes achievable by stimulation of S2.

The roots are split into the anterior and posterior components. The identity of the posterior root is confirmed by electrical stimulation and then a segment measuring ~20–40 mm in length is removed. When the S5 root has been identified, it is resected if no bladder response is obtained (39).

If a posterior rhizotomy is performed, stimulation can be applied to mixed sacral nerves in the sacral spinal canal extradurally, since the action potentials generated on the afferent axons do not reach the spinal cord. This has the advantage that the electrodes can be placed extradurally, reducing the risk of leakage of cerebrospinal fluid along the cables, and reducing the risk of breakage of the cables where they cross the dura. In addition, the extradural nerves are more robust than the intradural roots, being covered with epineurium derived from the dura, and require less dissection than the intradural roots; therefore, there is less risk of neuropraxia of the axons, which could otherwise lead to a delay in usage of the stimulator but not usually in permanent loss of function (1,41).

The benefits of a posterior rhizotomy include abolition of the neurogenic detrusor over activity, resulting in increased bladder capacity and compliance, reduced incontinence, and protection of the kidneys from ureteric reflux and hydronephrosis. The rhizotomy also reduces detrusor-sphincter dyssynergia, which improves urine flow, and prevents autonomic dysreflexia arising from distension or contraction of the bladder or bowel. In addition, a posterior rhizotomy improves implant-driven micturition. However, there are also drawbacks with a rhizotomy. They include abolition of reflex erection, reflex ejaculation, reflex defecation and sacral sensation, if present. Still, in many subjects with spinal lesions, these reflexes are not adequately functional, and function can be restored by other techniques (42).

The surgical technique for *extradural* implantation involves laminectomy of the first three pieces of the sacrum. It may also involve laminectomy of the L5 vertebra, depending on whether it is decided to implant electrodes on S2 roots (39). Extradural electrodes are used for patients in whom arachnoiditis makes separation of the sacral roots impossible. In some centers, however, extradural electrodes are used for all or nearly all patients.

After electrode implantation, the operation proceeded with closure of the dura, tunneling of the leads to a subcutaneous pocket in the flank, and closure of the skin. The patient is turned over and the leads are prepared for connection to the implantable stimulator.

At this time the leads are connected via an aseptic cable to an experimental stimulator. Prior to stimulation the bladder is filled with 200 mL saline using a transurethral filling catheter. The experimental stimulator consisted of two synchronized current sources with a common cathode. Pressure responses are elicited using pulse trains of 3–5 s duration; containing identical monophasic rectangular pulses delivered at a rate of 25 pulses $\cdot s^{-1}$. Stimulation is usually limited to the S3 and S4 ventral roots since they contain most of the motoneurons innervating the lower urinary tract.

After 15–20 min of experimental stimulation the leads are disconnected from the stimulator and the normal procedure is resumed with implantation stimulator.

A two-channel transurethral pressure catheter is used to measure intravesical and intraurethral pressure. The urethral pressure sensor is positioned at the level of the external sphincter such that in response to suprathreshold stimulation a maximal pressure response is measured. Pressures are sampled at 8 Hz, displayed on a monitor, and stored in a portable data logger (43).

All patients are followed up according to a fixed protocol. Urodynamic measurements are taken at 2 days, 15 days, 4 months, and 1 year after surgery and every 2–3 years thereafter. Renal ultrasound examination is performed every year. Stimulation is performed for the first time

Figure 28. BION microstimulator.

between days 8 and 14, depending on the level of the spinal cord lesion (33).

PUDENDAL NERVE STIMULATION FOR THE TREATMENT OF THE OVERACTIVE BLADDER (rf BION) (SEE FIGS. 28 AND 29)

Indications for use: The rf Bion system is still relatively new, and though no clear, established indications have been set so far, its activity on the pudendal nerve and inhibition of the detruser muscle makes it ideal for overactive bladder disorders.

Electrical stimulation of the pudendal nerve has been demonstrated to inhibit detrusor activity and chronic electrical stimulation may provide effective treatment for overactive bladder disorders (44). The hurdle to date has been the technical challenge of placing and maintaining an electrode near the pudendal nerve in humans; however, recent development of the BION has made chronic implantation feasible.

The BION is a small, self-contained microstimulator that can be injected directly adjacent to the pudendal nerve (see Fig. 28). The ischial spine is an excellent marker for the pudendal nerve as it re-enters the pelvis through Alcock's canal. This is a very consistent anatomical landmark in both men and women. Also, the implanted electrode is protected in this area by both the sacral tuberous and sacrospinous ligaments. Stimulation in this area activates afferent innervation over up to three sacral segments. Efferent stimulation also provides direct activation of the external urethral sphincter, the external anal sphincter, and the levator ani muscles, which may be of some benefit in bladder control. The external components of this neural prosthesis include a coil that is worn around the subject's hips and a controller that is worn around the shoulder or waist.

The technique chosen to implant the device is that of the transperineal pudendal block. This approach is minimally invasive and is well established. A special implant tool was

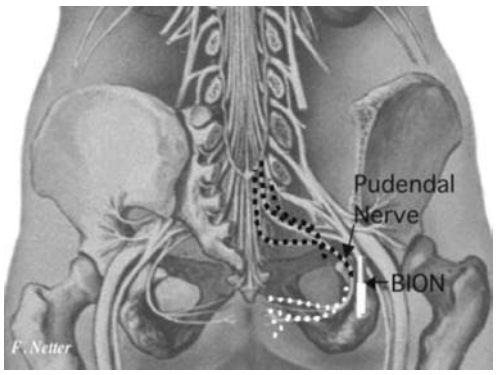


Figure 29. Placement of BION system near pudendal nerve.

devised to facilitate placement. The BION implantation technique was developed in cadavers. The optimum insertion location is 1.5 cm medial to the ischeal tuberosity using a vaginal finger to guide the implant toward the ischial spine where electrical stimulation of the pudendal nerve may be confirmed (see Fig. 29).

A percutaneous stimulation test (PST) was developed and proved to be a very effective way to assess acute changes in bladder volumes while stimulating the pudendal nerve. A baseline cystometrogram (CMG) was obtained followed by percutaneous pudendal nerve stimulation for 10 min with a repeat CMG.

The first implant was done on August 29, 2000. The BION was implanted under local anesthesia with intravenous sedation. Proper placement was verified by palpation and EMG activity. An intermittent stimulation mode of 5 s on 5 s off was used. Subjects returned 5–7 days later for activation, to distinguish between postoperative pain and potential stimulation pain. Subjects were followed up at 15, 30, and 45 days after activation. At each follow-up visit they underwent another cystometrogram and brought in a 72 h voiding diary. The results indicated a favorable response to maximum cystometric capacity throughout the study period. Diary entries verified improvement—incontinent episodes decreased by 65%, and both daytime and nighttime voids were decreased, as was pad use per day.

FUTURE DIRECTION OF THE THERAPY HARDWARE

The ongoing development of tools and hardware is driven by the desire to reduce the invasiveness of the implant and the likelihood of adverse events. Development efforts are concentrated on system components and tools that will allow implantation of the lead system through small incisions or percutaneous approaches. It is inevitable that the size of the neurostimulator will be reduced as future generations of the device are developed; more efficient power batteries and packaging will drive this aspect of development.

A rechargeable power battery may allow a smaller device. Although a smaller device would be welcomed, attaining this goal with a rechargeable battery is not seen as the best approach. A rechargeable neurostimulator would

require the patient frequently to recharge the unit; this would inconvenience the patient and could reduce patient compliance. Additionally, a rechargeable battery would be more expensive than a nonrechargeable one owing to the technology required and the additional equipment necessary for recharging. Furthermore, this would not eliminate the need for periodic replacement of the neurostimulator every 5–10 years. System components will be optimized for the therapy, to reduce the time needed for management of both implant and patient. The incorporation of microprocessors and implementation of features such as a battery gauge will provide additional operational information while decreasing the time needed to manage the patient.

Physicians will be able to analyze system use, lead status, and other parameters. The addition of sensing technology may provide an opportunity to create a closed-loop system that captures data to optimize both diagnosis and functioning.

Bilateral stimulation may provide more efficacious therapy. There is considerable interest in this approach, and it seems to be a probable avenue of research in the near future. However, any use of bilateral stimulation would have to justify the larger neurostimulator, the extra lead system, and the additional costs associated with this approach; at present, there is no scientific experience to support this approach.

Apart from a reduction in the size of the implanted device, enhanced physician control is the most likely development to occur in the foreseeable future. Graphics-based programming and control will simplify device programming; it will allow more complex features to be incorporated in the neurostimulator without adding undue complexity to the physician programmer. Management of patient data files will become easier as additional data-management features are added to the programmer; the physician will be able to obtain a patient-programming history and other patient-management data. It is conceivable that, in the not-so-distant future, the physician may be able to access patient-device data over the Internet, thus making unnecessary some clinic visits and allowing for remote follow-up of patients who are on holiday or have moved house.

Future devices may allow software loading in a noninvasive manner, to upgrade the device long after implantation. Such capability could be used to provide new therapy algorithms as well as new therapy waveforms.

The future will also bring enhanced test stimulation devices, which will provide improved fixation during the test stimulation period. The development of new leads is one such focus with the aim of allowing a longer test stimulation period without lead migration.

The future application of SNS is dependent on new clinical research. Pelvic disorders, such as pelvic pain and sexual dysfunction, appear likely to be the First areas of investigation; sacral anterior root stimulation for spinal cord injury may also provide a worthwhile avenue of enquiry. The development of these applications—or of any other, for that matter—will potentially require new waveforms and the development of new therapy algorithms. The future is as open as the availability of resources and the application of science allow (45).

BIBLIOGRAPHY

Cited References

1. Jezernik S, Craggs M, Grill WM, et al.Electrical stimulation for the treatment of bladder dysfunction: current status and future possibilities. Neurol Res 2002;24(5):413–430.
2. Galvani L. De virfbus electricitatis in motu musculari, commentaffus. De Bononiensi Scientarium et Artium Instituto Atque Academia. 1791;7:363–418.
3. Volta A. Letter to Sir Joseph Banks, March 20, 1800. On electricity excited by the mere contact of conducting substances of different kinds. Philos Trans R Soc London (Biol) 1800;90:403–431.
4. Duchenne GBA. De I'dlectrisation localisde et de son application & la physiologie, & la pathologie et b la therapeutique. Paris; 1855.
5. Duchenne GBA. Physiologie des moumments demonstrde b l'aide de l'experimentation electrique et de l'observation clinique, et applicable b I'6tude des paralysies et des deformations. Paris; 1867.
6. Chaffee EL, Light RE. A method for remote control of electrical stimulation of the nervous system. Yale J Biol Med 1934;7:83.
7. Glenn WWL, Phelps ML. Diaphragm pacing by electrical stimulation of the phrenic nerve. Neutosurgery 1985; 17:974–1044.
8. Glenn WWL, Mauro A, Longo E, et al.Remote stimulation of the heart by radiofrequency transmission. N Engl J Med 1959;261:948.
9. House WF. Cochlear implants. Ann Otol Rhinol Laryngol 1976;85(27):1–93.
10. Saxtorph MH. Strictura urethrae—Fistula petinei—Retentio urinae. Clinisk Chirurgi. Copenhagen: Gyldendalske Fodag; 1878.
11. Katona F, Benyo L, Lang J. Uber intraluminare elektrotherapie vor verschiedenen paralytischen zustlinden des gastrointestinalen tractes mit quadrangularstrom. Zentralbl Chir 1959;84:929.
12. Matona F. Stages otvegetative afferentation in reorganization of bladder control during electrotherapy. Urol Int 1975;30:192–203.
13. Schlote N, Tanagho EA. Electrical Stimulation of the lower urinary tract: historical overview. In: Jonas U, Grunewald V, editors. New Perspectives in sacral nerve stimulation. Dunitz; 2002. p 1–8.
14. McGuire WE. Response of the neurogenic bladder to vadous electrical stimuli [dissertation]. Department of Surgery, Bowman Gray School of Medicine; 1955.
15. Boyce WH, Lathem JE, Hunt LD. Research related to the development ofan artificial electrical stimulator for the paralyzed human bladder: a review. J Urology 1964;91:41–51.
16. Bradley WE, Chou SN, French LA. Further experience with the radio transmitter receiver unit for the neurogenic bladder. J. Neurosurg 1963;20:953–960.
17. Caldwell KPS. The electrical control of sphincter incompetence. Lancet 1963;2:174.
18. Fall M, Erlandson BE, Carlsson CA, Lindstr6m S. The effect of intravagnal electrical stimulation on the feline urethra and urinary bladder. Scand) Urol Nephrol (Supp) 1977;44: 19–30.
19. Lindstrim S, Fall M, Carlsson CA, Edandson BE. The neurophysiologcal basisaf bladder inhlbition in response to intravaginal electrical stimulation. Urology 1983;129:405–410.
20. McGuire EL, ZIang SC, Horwinski ER, Lytton B. Treatment of motor and sensory detrusor instability by electical stimulation. J Urol 1983;129:78–79.
21. Govier FE, Litwiller S, Nitti V, Kreder KJ, Jr., Rosenblatt P. Percutaneous afferent neuromodulation for the refractory

overactive bladder: results of a multicenter study. J Urol 165:1193, 2001.
22. van Balken MR, Vandoninck V, Messelink BJ, Vergunst H, Heesakkers JP, Debruyne FM, et al. Percutaneious tibial nerve stimulation as neuromodulative treatment of chronic pelvic pain. Eur Urol; 43:158, 2003.
23. van Balken, Michael R, Vergunst Henk, Bemelmans Bart LH. The use of Electrical Devices for the Treatment of Bladder Dysfunction: A Review of Methods. Urol September 2004;172(3):846–851.
24. Ingersoll EH, Jones LL, Hegre ES. Effect on urinary bladder of unilateral stimulation of pelvic nerves in the dog. Am Physiol 1957;189:167.
25. Hald T, Agrawal O, Mantrowitz A. Studies in stimulation of the bladder and its motor nerves. Surgery 1966;60:848–856.
26. Hald T, Meier W, Khalili A, et al. Clinical experience with a radio-linked bladder stimulator. J Urol 1967;97:73–78.
27. Friedman H, Nashold BS, Senechat R. Spinal cord stimulation and bladder function in normal and paraplegic animals. J Neurosurg 1972;36:430–437.
28. Jonas U, Heine JR, Tanagho EA. Studies on the feasibility of urinary bladder evacuation by direct spinal cord stimulation. 1. Parameters of most effective stimulation. Invest Urol 1975;13:142–150.
29. Jonas U, James LW, Tanagho EA. Spinal cord stimulation versus detrusor stimulation. A comparative study in six acute dogs. Invest Urol 1975;13:171–174.
30. Jonas U, Tanagho EA. Studies on the feasibility of urinary bladder evacuation by direct spinal cord stimulation. II. Poststimulus voiding: a way to overcome outflow resistance. Invest Urol 1975;13:151–153.
31. Nashold BS, Friedman H, Boyarsky S. Electrical activation of micturition by spinal cord stimulation. J Surg Res 1971;11:144–147.
32. Thirhoff JW, Bazeed MA, Schmidt RA, et al. Regional topography of spinal cord neurons innervating pelvic floor muscles and bladder neck in the dog: a study by combined horseradish peroxidase histochemistry and autoradiography. Utol Int 1982;37:110–120.
33. Tanagho EA, Schmidt RA. Bladder pacemaker: scientific basis and clinical future. Urology 1982;20:614–619.
34. Schmidt RA, Bruschini H, Tanagho EA. Sacral root stimulation in controlled micturition: peripheral somatic neurotomy and stimulated voiding. Invest Urol 1979;17:130–134.
35. Probst M, Piechota HA, Hohenfeliner M, et al.Neurostimulation for bladder evacuation: is sacral root stimulation a substitute for microstimulation? Br J Urol 1997;79:554–566.
36. Bosch JLHR, Groen J. Sacral nerve neuromodulation in the treatment of patients with refractory motor urge incontinence: long-term results of a prospective longitudinal study. J Urol 2000;163:1219.
37. Spinelli M, Giardiello G, Gerber M, Arduini A, Van Den Hombergh U, Malaguti S. New Sacral Neuromodulation Lead For Percutaneous Implantation Using Local Anesthesia: Description And First Experience. J Urol 2003;170(5): 1905–1907.
38. Brindley GS. The first 500 patients with sacral anterior root stimulator implants: general description. Paraplegia 1994; 32:795–805.
39. Egon G, Barat M, Colombel P, et al.Implantation of anterior sacral root stimulators combined with posterior sacral rhizotomy in spinal injury patients. World J Urol 1998;16:342–349.
40. Brindley GS, Polkey CE, Ruston DN. Sacral anterior root stimulators of bladder control in paraplegia. Paraplegia 1982;28:365–381.
41. Brindley GS, Polkey CE, Rushton DN, Cardozo L. Sacral anterior root stimulators for bladder control in paraplegia: The first 50 cases. J Neurol Neurosurg Psychiat 1986;49: 1104–1114.
42. Rijkhoff N. Neuroprostheses to treat neurogenic bladder dysfunction: current status and future perspectives. Childs Nerv Syst 2004 Feb; 20(2): 75–86.
43. Rijkhoff N, Wijkstra H, Kerrebroeck P, et al.Selective detrusor activation by sacral ventral nerve-root stimulation: results of intraoperative testing in humans during implantation of a Finetech-Brindley system. World J Urol 1998;16: 337–341.
44. Vodusçek DB, Light KJ, Libby JM. Detrusor inhibition induced by stimulation of pudendal nerve afferents. Neurourol Urodyn 1986;5:381.
45. Gerber M, Swoyer J, Tronnes C. Hardware: development and function. New Perspectives in sacral nerve stimulation. In: Jonas U, Grunewald V, editors. Dunitz: 2002. p 81–88.

See also BIOTELEMETRY; FUNCTIONAL ELECTRICAL STIMULATION; TRANSCUTANEOUS ELECTRICAL NERVE STIMULATION (TENS).

BLIND AND VISUALLY IMPAIRED, ASSISTIVE TECHNOLOGY FOR

ANDREW Y. J. SZETO
San Diego State University
San Diego, California

INTRODUCTION

Severe visual impairment represents one of the most serious sensory deficits that a human being can have. When this sensory input channel is so impaired that little useful information can pass through it, assistive devices that utilize alternative sensory input channels are often necessary. Familiar examples include the use of Braille and the white cane, respectively, for reading and obstacle avoidance by persons who are blind. Both of these assistive devices provide environmental information to the user via the sense of touch. Other assistive devices provide environmental feedback via the sense of hearing.

In the material that follows, examples of available assistive technology and promising new assistive technology under development for persons who are blind or severely visually impaired are presented. This article begins with an overview of the prevalence and impairments associated with blindness impairments and follows with an examination of reading aids, independent living aids, and mobility aids. The article concludes with a brief look at kinds of assistive technology likely to be available in the near future for persons with severe visual impairments.

The term blindness has many connotations and is difficult to define precisely. To many people, blindness refers to the complete loss of vision with no perception of light. The U.S. government, however, defines blindness as the best corrected visual acuity of 20/200 or worse in the better seeing eye. The acuity designation 20/200 means that a vision impaired person is able to see at a distance of 20 ft (6.09 m) what a person with normal visual acuity is able see at 200 ft (60.96 m). Low vision is defined as the

Table 1. Prevalence of Blindness and Low Vision Among Adults 40 Years and Older in the United States[a]

	Blindness		Low Vision		All Vision Impaired	
Age, Years	Persons	%	Persons	%	Persons	%
40–49	51,000	0.1	80,000	0.2	131,000	0.3
50–59	45,000	0.1	102,000	0.3	147,000	0.4
60–69	59,000	0.3	176,000	0.9	235,000	1.2
70–79	134,000	0.8	471,000	3.0	605,000	3.8
>80	648,000	7.0	1,532,000	16.7	2,180,000	23.7
Total	*937,000*	*0.8*	*2,361,000*	*2.0*	*3,298,000*	*2.7*

[a]Abstracted from Ref. 3 Arch. Ophthalmol. Vol. 122, April 2004.

best corrected visual acuity that is worse than 20/40 in the better seeing eye. People with extreme tunnel vision (a visual field that subtends an angle $>20^\circ$ regardless of the acuity within that visual angle) also are classified as being legally blind and thus qualify for certain disability benefits.

It is important to realize that a great majority ($\sim$70–80%) of people with severe impairments has some degree of usable vision (1,2). The severity of vision loss can vary widely and result in equally varying degrees of functional impairment. Although the degree of impairment may differ from one person to another, people who are blind or have low vision experience the common frustration of not being to see well enough to perform common everyday tasks.

The prevalence of blindness and low vision among adults 40 years and older is given in Table 1. According to the National Eye Institute (2), a component of the National Institutes of Health in the United States Department of Health and Human Services, the leading causes of vision impairment and blindness are primarily age-related eye diseases. These include age-related macular degeneration, cataract, diabetic retinopathy, and glaucoma. The 2000 census data revealed >5 million people of all ages in America have visual impairments severe enough to significantly interfere with their daily activities.

CONSEQUENCES OF SEVERE VISUAL IMPAIRMENTS

The two major difficulties faced by persons who are blind or severely visually impaired are access to reading material and independent travel or mobility. Simple-to-sophisticated technology has been used in a variety of assistive devices to help overcome these problems. The term reading is used in this context to include access to all material printed on paper or electronically. Reading material can include text, pictures, drawing, tables, maps, food labels, signs, mathematical equations, and graphical symbols. Safe and independent mobility is used to encompass both obstacle avoidance and navigation. For safe and independent mobility, the first concern is avoiding obstacles, such as curbs, chairs, low hanging branches, and platform drop-offs. After the sight impaired traveler has gained an awareness of the basic spatial relationships between objects within the travel environment, their needs wayfinding or navigational assistance, which involves knowing one's position, one's heading with respect to the intended destination, and a suitable path to reach it.

LOW VISION READING AIDS

People with low vision significantly out number those who are totally without sight (Table 1). Hence, the consumer market for low vision aids is much larger than the one dedicated to people with zero vision. The technology used in low vision aids is rather straightforward and the technologically used is relatively mature. Hence, only a brief overview of such assistive devices will be presented before discussing the more challenging issues faced by persons with zero useful vision. For readers desiring detailed product information about low vision aids, a search of the Internet using the term low vision aids will yield a bounty of pictures, product specifications, and purchasing information.

All low vision aids aim to maximize an individual's residual vision to its fullest. Low vision aids can be categorized as optical, nonoptical, and electronic. Optical aids include handheld magnifying glasses, telescopes mounted on eyeglass frames, and even microscope lenses. Nonoptical aids include enlarged high contrast print and high intensity lamps.

Electronic low vision aids represent the highest level in terms of cost, complexity, and performance. They include electronic video magnifiers that project printed material on a closed circuit monitor, regular television, or computer screen. Electronic video magnifiers can maximize readability of the written material by providing a wide range of magnification, brightness, contrast, type of fonts, and foreground and background colors. A good example of a modern closed circuit TV type of electronic low vision aid is the Optelec Traveller (Fig. 1). This portable video

Figure 1. This portable video magnifier has a built-in 6 in. (15.24 cm) color screen and can magnify text and pictures up to 16 times. (Courtesy of Optelec International, New York.)

Figure 2. Closed-circuit television with computer based text-to-speech output, a talking computer.

magnifier has a built-in 6 in. (15.24 cm) color screen and can magnify text and pictures up to 16 times and more if its video signal is sent to a television set.

People with tunnel vision or central blind spots due to macular degeneration often find it difficult and tiring to read an entire computer screen. For such individuals, the advent of the talking computer (Fig. 2) represented a major technological breakthrough. The capability and flexibility of such a computer or reading machine addressed many of their needs as well as the needs of persons without any useful vision.

READINGS AIDS FOR THE BLIND

For persons with essentially zero useful vision, the tactile sense has been utilized as an alternative sensory input channel for reading. One of the oldest reading substitutes for the blind is Braille, a six dot matrix code that Louise Braille adapted in 1824 for use by blind persons to read written text. The standard Braille cell consists of two columns and three rows of dots separated by 2.3 mm with 4.1 mm separating adjacent cells. Each Braille cell occupies a rectangular area of 4.3×8.6 mm and can represent 2^6-1 (or 63) possible symbols within that areas. Grade I Braille maps each cell a one-to-one basis to each letter of the alphabet, basic punctuation marks, and simple abbreviations so that Grade I Braille has an informational density of approximately 1 bit per 6 mm^2 of surface area. For greater informational compactness and faster reading rates, Grade II Braille uses combinations of dots to represent contractions, frequently used words, prefixes, and suffices. Grade III Braille is even more compact and affords the highest reading rates, but very few people ever master it. The largest proportion of Braille literature is produced at the Grade II level, which can be read at up to 200 words per minute (4) by those proficient in Braille. Braille is a unique reading aid that not only gives blind persons access to printed material but also provides them with a writing medium.

Despite Braille's unique place as a complete writing system that is spatially distributed and retains many advantages of a printed page, Braille is a specialized code that only a small percentage of blind individuals learn to use. This is especially true for persons who become blind after the age of 15 years. Given the difficulty of mastering Braille, the lack of up-to-date Braille printed material, and advances in alternate technologies such as electronic reading machines, many blind individuals choose to not bother with Braille.

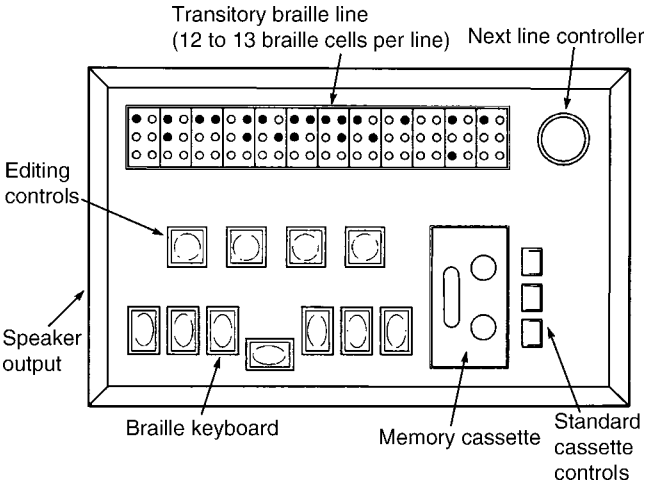


Figure 3. Portable refreshable Braille reader that can playback or store messages using a cassette recorder. The reader has a single-line tactile display, a Braille keyboard, and a tape cassette for data storage and recall. (Picture taken from Fig. 2.8 of Ref. 5.)

Other disadvantages of Braille printed material include the cost to produce it, store it, and maintain it. Embossable Braille paper is not only bulky, heavy, and expensive, the pattern of raised dots (laboriously and noisily impressed into the paper) is fragile and short lived. Assistive technologies such as portable Braille readers (Fig. 3) have mitigated some of the inconveniences associated with Braille (5), but these electronic Braille readers–recorders often do not display the two-dimensional (2D) information embedded in graphs, tables, and mathematical formulas. The single and dual line tactile displays found in most portable readers also makes the rapid search for content via headings very difficult.

Refreshable Braille readers can be used as a computer interface for accessing information on the computer screen. Some full-sized electronic Braille displays are 80 cells long and cost upward of $10,000. The dots in these transient Braille displays are produced by pins raised and lowered (refreshed) to form Braille characters. Refreshable Braille readers allow users to access any portion of the screen information via specialized control buttons and status Braille cells. Tactually distinguishable arrow keys offer screen cursor control while extra status cells provide additional information about text-attributes or line and colon positions.

Refreshable Braille displays are especially useful for deaf blind individuals and users working with computer programming languages. For example, the Braille Voyager 44 (Fig. 4), made by F.J. Tieman BV, has a 44 cell Braille display, and 5 thumb keys for screen navigation. Using its built-in macro program, USB connection, and any screen reader, the Voyager enables a user to access many features of the Windows operating system.

Figure 4. The Braille Voyager 44 made by Manufacturer: F.J. Tieman BV. It has a 44 cell Braille display and 5 navigation keys.

Despite Braille's many drawbacks and limited popularity, its long history, status as the only complete writing and reading system for the blind, and tenacity of advocates like the American Federation for the Blind combine to keep Braille viable as an informational medium. Nonprofit groups like the Braille Institute produce millions of pages of Braille each year for business, schools, government agencies and individuals across the nation. They sell recreational reading material in Braille to both children and adults and provide low cost transcription, embossing, and tactile graphic services.

For the majority of blind persons who do not know Braille, reading material converted into the audio format (aka talking books) and played back on variable speed tape recorders have proven to be popular and convenient to use. To overcome spoken speech's inherently slower reading rate, variable speed tape recorders with special electronic circuits that compensate for the pitch change during high speed playback (1.5–3 times normal speed) can be used. Obtaining reading material in audio form for playback on such recorders also has become more convenient as vendors make downloading of electronic text and audio files available to their subscribers (6).

Although audio books are popular for persons with severe visual impairments, this approach does not work for reading the newspaper, daily mail, memoranda, cookbooks, technical reports, handwritten notes, and common everyday correspondence, such as utility bills and bank statements. Before the advent of a reading machine, which has now become part of a general purpose talking computer, persons with no useful vision relied on human readers with its attendant inconvenience, loss of independence, and lack of privacy.

For severely sight impaired individuals and even those who know Braille, the power, convenience, and versatility of a reading machine, also known as a talking computer, have made it the preferred method of accessing most reading material. First marketed in the early 1980s, reading machines of today are affordable, compact, and can reliably and rapidly convert alphanumeric text into synthetic speech. In addition to a synthetic voice that reads aloud the actual text, the talking computer or reading machine also provides auditory feedback of cursor location and navigational commands.

A talking computer or dedicated reading machine contains artificial intelligence that converts alphanumeric text into spoken speech. The multistep process begins with an optical device that scans the text of a printed document or web page and, using optical character recognition, converts that alphanumeric text string into prefixes, suffixes, and root words (Fig. 5).

The process through which the text string is converted into speech output is somewhat complex and undergoing refinement. The clarity and naturalness of the voice output depend on the text-to-speech technique employed. In general, clearer and more natural costlier sounding speech requires more memory and greater processing power and is thus more expensive.

After the written material has been converted into a text string by optical character recognition software, one of

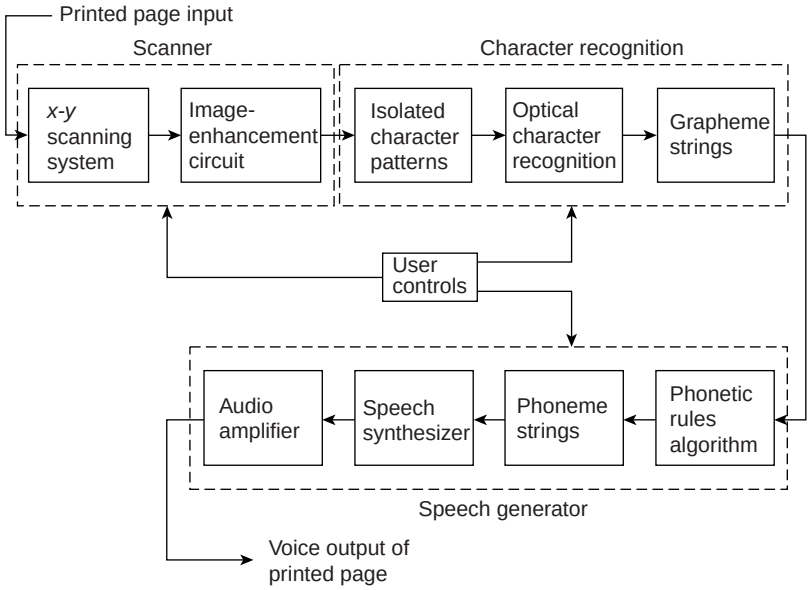


Figure 5. Functional components of a reading machine. (Taken from Fig. 2.18 of Ref. 5.)

three basic methods can be employed to convert the string into speech sounds. The first method is called whole word look-up. It produces the most intelligible, life-like speech, but it is the most memory intensive even for modest sized vocabularies. Despite steady advances in low cost, high density memory chips, whole-word-look-up tends to be prohibitively expensive or the vocabulary is limited (7).

A less memory intensive approach is the letter-to-sound conversion in which a synthetic sound processor divides the text string into basic letter groups and then follows certain pronunciation rules for the creation of speech. Many languages (especially American English) are replete with numerous exceptions to the usual rules of pronunciation. Hence, the quality of the speech output using letter-to-sound conversion depends on the sophistication of the rules and the number of exceptions employed (7,8).

The third method of converting text into speech is called morphonemic text-to-speech conversion. This approach relies on prestored combination of morphemes (basic units of language such as prefixes, suffixes, and roots) and their corresponding speech sounds. Some 8000 morphemes can generate $\sim 95\%$ of the English words (8,9) so this approach avoids the memory demands of the whole word look-up approach. Morphonemic based speech generation generally yields synthetic speech output that is more intelligible than the letter to speech approach, but is more demanding computationally. Continuing advances in technology have now made single chip text to speech converters powerful, capable, and affordable in consumer electronics (10).

A blind individual using a computer running a text-to-speech program can now hear what is on the screen and use cursor keys to select a specific part of the screen to read. Equipped with such a computer, high speed connection to the Internet, and a modern reading machine, sight impaired individuals now have wide access to news, e-mail, voice messaging, and Internet's vast repository of information. These powerful information technologies have reduced the social isolation formerly felt by blind persons while also broadening their employment opportunities.

One example of how recent technological advances are improving access to reading materials is the Spoken Interface that Apple Computer unveiled at the 2005 Annual Technology & Persons with Disabilities Conference held in Los Angeles. Because Spoken Interface is a screen reader that is fully integrated into Apple's operating system, assistive technology developers should be able to set up easy inter-operability between their software and the operating platform with little additional modifications.

Another example of a low cost, user friendly, and powerful text-to-speech software is the TextAloud MP3 by Nextup Technologies (http://www.nextuptech.com/about.html). This software converts any text into natural sounding speech or into MP3 files for downloading and later playback on portable electronic devices (e.g., MP3 players, pocket PCs, and portable data assistants).

MANDATED WEB ACCESSIBILITY

With so much information available on the Internet and the blind people's increasing dependence on it, the United States government included web accessibility in its 1998 amendment of the Rehabilitation Act (11). Section 508 of this law requires that when Federal agencies develop, procure, maintain, or use electronic information technology, they must ensure that this technology offers comparable access to Federal employees who have disabilities. Although the scope of Section 508 is limited to the Federal sector, these requirements have gradually spread to the private sector, especially to large corporations that deal frequently with the Federal government.

The accessibility requirements of Section 508 are reflected in several guidelines, including as the Web Content Accessibility Guidelines (WCAG) from the World Wide Web Consortium (W3C). The WCAG recommendations, which are updated periodically, include implementing standardized style sheets instead of custom HTML tags and offering closed-captioning and transcripts of multimedia presentations. Other recommendations for making a web site compliant (12) include the following: provide text alternates to images; make meaning independent of color; identify language changes; make pages style sheet independent; update equivalents for dynamic content; include redundant text links for server-side image maps; use client-side image maps when possible; put row and column headers in data tables; associate all data cells with header cells; title all frames; make the site script independent.

An assortment of adaptive hardware and software can be effectively utilized once a web site satisfies the WCAG recommendations (13). Persons with low vision can change their browser settings or use screen magnifiers. Internet users who are blind or have very limited vision can use text-based Web browsers with voice-synthesized screen readers, audio browsers, or refreshable Braille displays to read and interact with the Web.

Recent efforts to increase internet's compatibility with assistive technologies used by sight impaired persons include the development and implementation of search engines that read aloud their results using male and female voices. Some websites offer speech-synthesized renditions of articles from news organizations like BBC, Reuters, and the New York Times (14).

While internet accessibility by persons with severe visual impairments is improving, a number of problems and challenges remain. Screen readers or Braille keyboards that blind people use to navigate the Internet cannot scan or render graphical elements into a readable format. Spam, security checks, popup ads, and other things that slow down a sighted person's Web searches are even worse impediments for those with severe visual impairments using assistive technology.

INDEPENDENT LIVING AIDS

Because blindness and severe visual impairments are so pervasive in their impact, numerous and relatively low cost assistive devices have been developed to make non-reading activities of daily living (ADL) easier. In general, these ADL devices rely on the users' auditory or tactile sense for their operation.

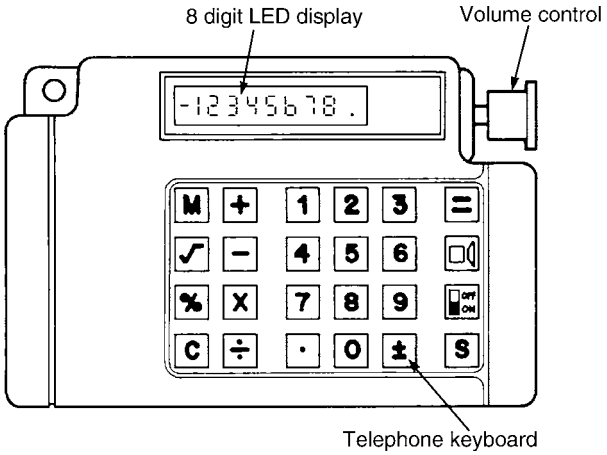


Figure 6. A talking calculator with a female voice speaks the individual digits or whole integers. Its large 8 digit LCD readout is ~0.6 in. (1.52 cm) high. The calculator can add, subtract, divide, multiply and calculate percentages. (Reproduced from Ref. 5.)

A quick check of electronic catalogs on the Internet shows that many types of independent living aids are available. For example, special clocks and timers that give both voice and vibratory alarms are available in various sizes and features. Other assistive devices for ADL include talking wrist watches, push-button padlocks, special money holders, Braille embossed large push button phones, and jumbo sized playing cards embossed with Braille. Personal care items for the blind include talking bathroom scales, thermometers, glucose monitors, blood pressure gauges, and prescription medicine organizers. Educational aids that facilitate note taking, calculating, searching, printing, and organizing information include talking calculators (Fig. 6), pen-like handheld scanner for storing text, letter writing guides with raised lines, Braille metal guides and styluses, and signature guides.

MOBILITY AIDS

For persons with severe visual impairments, the advent of powerful and affordable reading machines and the vast amount information (already in electronic form) on the internet, the problem of access to reading materials has been significantly ameliorated. In contrast, their other major problem (the ability to travel safely, comfortably, gracefully, and independently through the environment) has only been partially solved.

Despite years of effort and some major advances in technology, there is no widely accepted electronic travel aid (ETA). Most blind individuals rely on the sighted human guide, a guide dog, and the familiar white cane. The human sighted guide offers companionship, intelligence, wayfinding capability, route recall, and adaptability. Unfortunately, human guides are not always available, and their very presence constitutes a lack of independence. A guide dog or animal guide has been popular, but not every blind person can independently care for a living animal nor afford the cost of its care. In some social situations, a guide dog can be awkward or unacceptable. The white cane, which is both a tool and a symbol for the blind, can alert sight-impaired travelers to obstacles in their path, but only those at ground level and < 5 ft. (1.5 m) away. Above ground obstacles and especially those at head height remain a source of apprehension and danger for travelers depending on just the white cane.

To understand why decades of research and development efforts have not yielded an efficacious and widely accepted electronic travel aid, one needs to realize that mobility aids must deal with a very different set of constraints and inputs than do reading aids. An identification error made by reading aids results only in misinformation, mispronunciation, or inconvenience. In contrast, a failed detection of an obstacle or step-down or a missed landmark can lead to confusion, frustration, apprehension, and physical injury.

Another major difference between a mobility aid and a reading aid lies in their operating milieu. Mobility aids must detect and analyze unconstrained, long range, and highly variable environmental inputs, that is, obstacles of differing sizes, textures, and shapes distributed over a 180° wide area. In contrast reading machines must identify and convert into intelligible speech inputs that are often well defined and short ranged, for example, high contrast printed alphanumeric symbols and punctuation marks (15).

To further complicate matters, users of reading aids often have the luxury of focusing all or most of their attention on the task at hand: interpreting the output of the reading aid. Users of mobility aids, however, must divide their attention among several demanding tasks associated with traveling, such as avoiding obstacles, listening to environmental cues, monitoring their physical location, recalling the memorized route, and interpreting the auditory or tactile cues from their mobility aid. Given these challenges, today's mobility aids represent a much less satisfactory solution (in comparison to available reading aids) to the problem of independent and safe mobility for persons with severe visual impairments.

THE IDEAL MOBILITY AID

Before examining the capabilities of currently available mobility aids, it is desirable to enumerate the fundamental features of an ideal electronic travel or mobility aid (Table 2) (16–18). The first three items of an ideal mobility aid can be categorized as nearby obstacle avoidance; features 4–7 fall under the category of navigational guidance or wayfinding; and features 8–10 represent good ergonomic design or user friendliness.

CONVENTIONAL ELECTRONIC TRAVEL AIDS

Standard or conventional electronic travel aids detect nearby obstacles, but provide no wayfinding assistance. Obstacle detection entails the transmission of some sort of energy into the surrounding space and the detection of the reflections. After analyzing the reflected signals, the ETA warns the traveler of possible obstacles using either auditory feedback or tactile feedback.

Table 2. The Ideal Mobility Aid

	Capabilities and Features	Description
Feature No. 1	Obstacle detection	Detect nearby obstacles that are ahead, at head level, and at ground level and indicate their approximate locations and distances without causing sensory overload.
Feature No. 2	Warn of impending Obstacles	Reliably locate and warn of impending potholes, low obstacles, step-downs and step-ups.
Feature No. 3	Guidance around obstacles	Guide the traveler around impending obstacles.
Feature No. 4	Ergonomically designed	Offer voice and/or tactile feedback of traveler's present location. Capable of voice input operation and/or have tactually distinct push buttons
Feature No. 5	Wayfinding	Able to monitor the traveler's present location and indicate the direction toward the destination
Feature No. 6	Route recall	Be able to remember a previous route and warn of changes in the environment due to construction or other blockages
Feature No. 7	Operational flexibility	Reliably function in a variety of settings, that is, outdoors, indoors, stairways, elevators, and cluttered open spaces
Feature No. 8	User friendliness	Be portable, rugged, fail-safe, and affordable for a blind user
Feature No. 9	Cosmesis	Be perceived by potential users as cosmetically acceptable and comfortable to use in terms of size, styling, obtrusiveness, and attractiveness
Feature No. 10	Good battery life	Have rechargeable batteries that can last for at least 6 h per charge

The LASER CANE (Fig. 7) is one of the few conventional ETAs that can serve as a stand-alone, primary travel aid because it has obstacle detection (features 1–3 of Table 2) and is reasonably user friendly and cosmetic (features 8–10). The laser cane's shaft houses three narrow-beam lasers; the lasers scan upward, forward, and downward. Reflections from objects in these zones are detected by three optical receivers also housed in the shaft. The UP channel monitors head level obstacles and causes high pitched beeps to be emitted. The FORWARD channel monitors objects located 4–10 ft. (1.21–3.01 m) ahead of the cane's tip and produces warning signals in the form of either vibrations in the handle of the cane or a medium (1600 Hz) audio tone. Obstacles encountered by the DOWN channel produce a low frequency (200 Hz) warning tone (19). Because the laser cane is swept through an arc $\sim$3–4 ft. (0.91–1.21 m) wide in the direction of the intended path (in a manner similar to standard long cane usage), the laser cane augments the auditory and tactile feedback of an ordinary white cane by detecting objects at greater distances and, most importantly, head level obstructions.

The laser cane's main drawbacks include it being somewhat costly and fragile. It also cannot monitor the traveler's geographic location nor guide the traveler toward the intended destination (features 5 and 6). Field tests and consumer feedback revealed that laser obstacle detection can be highly variable because certain surfaces and objects reflect laser light better than others. For example, the laser beam mostly passes through glass so that the laser cane may miss glass doors or large glass windows ahead.

Although the laser cane is imperfect, it has one major advantage as an ETA; It is failsafe. Should its batteries run down or its electronics malfunction, the laser cane can still serve as a standard long cane (20) and thus still be useful to the traveler.

Another commercially available electronic travel aid is the Sonic Guide, an eyeglass frame equipped with one ultrasonic transmitter and two receivers embedded in

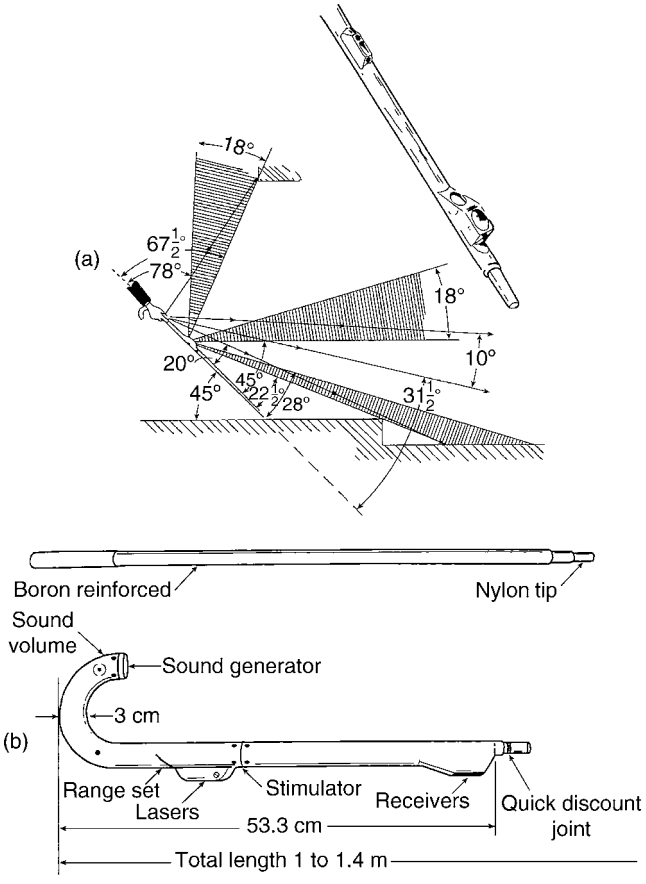


Figure 7. The laser cane projects three narrow beams of laser light. If any of the beams (up, forward, and downward) encounter an object and is reflected back to the receivers in the cane's shaft, a tactile or auditory warning is generated. (Reproduced from Ref. 19.)

the nose piece (21). The pulsed ultrasonic beam radiates through a forward solid angle of $\sim 100^\circ$. Objects in the environment reflect ultrasound back to the two receivers with time delays proportional to their distance and angle with respect to the wearer's head. The wearer is given awareness of his surroundings via binaural auditory feedback of the reflected signals, recreating the experience of echolocation as found in bats or dolphins. An object's distance is displayed in terms of frequencies proportional to the object's distance from the user. The azimuth of an object relative to the user's head is displayed via the relative intensity of tones sent to the ears (stereoscopic aural imaging). As a result, the binaural sounds heard by the user changes as he moves or turns his head.

To circumvent Sonic Guide's tendency to interfere with normal hearing, Kuc (22) investigated the utility of using vibrotactile feedback via a pair of sonar transceivers and vibrators worn on the wrists. Being on opposite sides of the body, the dual sonar transceivers offered better left–right obstacle discrimination than could a single sonar unit embedded in the nose piece of the eyeglasses. The wrist mounted pager-like device vibrated at a frequency inversely related to the reflecting object's distance from that side of the body.

Unfortunately, neither the original eyeglass frame based Sonic Guide nor the wrist worn sonar guide can serve as a stand-alone travel aid because neither can detect impending step-ups, step-downs, or other tripping hazards in the pathway. Other user comments about the Sonic Guide include interference with normal hearing, sensory overload, and difficulty in combining the aid's feedback with other important environmental cues such as the sound of traffic at street intersections, tactile feedback from a white cane, or the subtle pull of a guide dog.

In contrast to Sonic Guide's rich auditory feedback, the Mowat Sensor implements the design philosophy that simpler is better. The Mowat Sensor is a handheld ultrasonic flash light that acts like a clear path detector. It measures $6 \times 2 \times 1$ in. ($15 \times 5 \times 2.5$ cm), weighs 6.5 oz (184.2 g), can be easily carried in a pocket or purse, and is manufactured by Pulse Data International Ltd. of New Zealand and Australia.

The Mowat device emits a pulsed elliptical ultrasonic beam $\sim 15^\circ$ wide by 30° high, a beam pattern that should detect doorway sized openings located some 6 ft. (1.8 m) away. Reflections from objects in the beam pattern cause the Mowat to produce vibrations that are inversely proportional to the object's distance from the detector. As the traveler points at and gets closer to the object, the Mowat vibrates faster and faster. As the traveler aims moves away from that object, the vibrations slow and then cease. Objects outside of Mowat's beam pattern produce no vibrations.

The Sonic Guide, Mowat Sensor, and their various derivatives share similarities while representing two divergent design philosophies. They all employ ultrasound instead of laser light to detect nearby obstacles. None of them can detect tripping hazards, such as impending step-ups, step-downs, uneven concrete walkways, or small low obstacles in the path of travel so they cannot serve as a stand-alone travel aid. The Mowat sensor scans a small portion of the environment, displays limited data from that region, and offers easily interpreted vibratory information to the user. Alternatively, the Binaural Sonic Guide sends a broad sonic beam into much of the traveler's forward environment, displays large amounts of environmental information, and leaves it up to the user to select which portion of the auditory feedback to monitor and which to ignore.

While similar in concept, obstacle detection via ultrasound and obstacle detection via laser light interact with the environment differently. For example, hard vertical surfaces and glossy painted surfaces reflect sound and light very well so they tend to be detected by both methods at greater distances than oblique surfaces or dark cloth covered soft furnishings. Transparent glass, however, reflects sound very well, but laser light very poorly. Hence an ultrasonic beam would readily note the presence of a glass door whereas laser light could miss it entirely. Sonar based ETAs, however, are susceptible to spurious sources of ultrasound such as squealing air breaks on buses. Such sources and even heavy precipitation can cause the sonar sensor to signal the presence of a phantom obstacle or produce unreliable feedback. Furthermore, because all ETAs display environmental information via the sense of touch or hearing, severe environmental noise and wearing gloves or ear muffs can reduce a user's ability to monitor an ETAs feedback signals.

Other drawbacks of conventional electronic travel aids include the lack of navigational guidance (features 5 and 6 of Table 2), thus limiting the blind traveler to familiar places or necessitating directional guidance from a sighted guide until they have memorized the route. Furthermore, conventional ETAs often require the user to actively scan the environment and interpret the auditory and tactile feedback from the aids. These somewhat burdensome tasks require conscious effort and can slow walking speed.

INTELLIGENT ELECTRONIC TRAVEL AIDS

Recent advances in technology have sparked renewed efforts to develop mobility aids that address some of the aforementioned drawbacks. One promising intelligent electronic travel aid, under development at the University of Michigan Mobile Robotics Laboratory, is the GuideCane (23). The GuideCane (Fig. 8) is a semiautonomous robotic guide that improves user friendliness by obviating the burden of constant scanning while also guiding the traveler around obstacles, not merely detecting them. It consists of a self-propelled and servocontrolled mobile platform connected to a cane. An array of 10 ultrasonic sensors is mounted on the small platform. The sensors emit slightly overlapping signals to detect ground-level obstacles over a 120° arc ahead of the platform. The sonar units, made by Polaroid Corporation, emit short bursts of ultrasound and then uses the time of flight of the reflections to gauge the distance to the object. The sonar has a maximum range of 30 ft. (10 m) and an accuracy of $\sim 0.5\%$ (24).

When walking with the GuideCane, the user indicates his intended direction of travel via a thumb-operated minijoystick mounted at the end of a cane attached to the

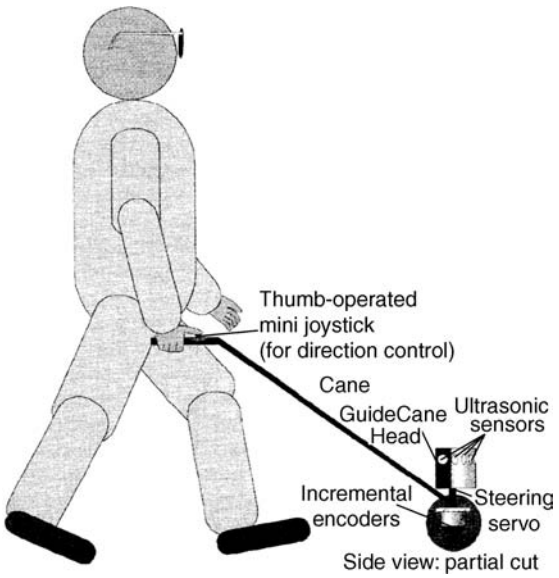


Figure 8. The GuideCane functions somewhat like a robotic guide dog. It is able to scan the environment and steer around obstacles by using its ultrasonic sensors, steering servomotors, and on-board computer to keeps track of nearby obstacles and the intended path of travel. (Reprinted from figure on p. 435 of Ref. 23.)

platform. The mobile platform maintains a map of its immediate surroundings and self-propels along the indicated direction of travel until it detects an obstacle at which time the robotic guide steers itself around it. The blind traveler senses the GuideCane's change of direction and follows it accordingly.

Like the Laser cane, the GuideCane can function as a stand-alone travel aid because it gives advance warning of impending step-downs and tripping hazards. Its bank of 10 ultrasonic detectors and ability to navigate around detected obstacles make the GuideCane easier and less mentally taxing to use than the Laser Cane. To address the wayfinding needs of the blind traveler, efforts are underway to add GPS capability, routing software and area maps to the GuideCane. The drawbacks of the wheel mounted GuideCane, however, include its size and weight and its inability to detect head height objects.

NAVIGATIONAL NEEDS

Electronic travel aids like those described above are becoming proficient at detecting and enabling the traveler to avoid obstacles and other potential hazards. Avoiding obstacles, however, represents only a partial solution to a blind person's mobility problem. Many visually impaired or blind travelers hesitate to visit unfamiliar places because they fear encountering an emergency or possibly getting lost. Their freedom of travel is hampered by having to pre-plan their initial trip to a new place or needing to enlist the help of a sighted person.

Furthermore, blind pedestrians, even those with training in orientation and mobility, often experience difficulty in unfamiliar areas and areas with free flowing traffic, such as parking lots, open spaces, shopping malls, bus terminals, school campuses, and roadways or sidewalks under construction. They also have difficulty crossing nonorthogonal, multiway traffic intersections (25). Conventional traffic signals combined with audible pedestrian traffic signals have proven somewhat helpful in reducing the pedestrian accident rates at intersections (26–28), but audible traffic signals offer guidance only at traffic intersections and not other important landmarks.

One proposed solution for meeting the wayfinding needs of blind travelers is the Talking Sign, a remote infrared signage technology that has been under development and testing at The Smith-Kettlewell Eye Research Institute in San Francisco, CA (29,30). The Talking Signs system consists of strategically located modules that transmit environmental speech messages to small, hand held receivers carried by blind travelers (Fig. 9). The repeating and directionally selective voice messages are transmitted to the receiver by infrared (IR) light (940 nm, 25 kHz). Guided by these orientation aids, blind travelers can know their present location and move in the direction from which the desired message, for example, Corner of Front Street and Main Street, is being broadcasted, thus finding their way without having to remember the precise route.

The Talking Sign and other permanently mounted voice output devices, however, require standardization, costly retrofitting of existing buildings, and the possession of a suitable transceiver to detect or activate the installed devices. Retrofitting buildings with such devices is not cost effective due to their inherent inflexibility and the need for many users to justify the implementation costs. What's especially frustrating for persons with severe visual impairments is that talking signs may not reflect their travel patterns or be available at unfamiliar locations and wide open spaces. To be truly useful, talking signs would have to be almost ubiquitous and universally adopted.

GPS NAVIGATIONAL AIDS

In addition to obstacle avoidance, the ideal navigational aid also must address two other key aspects of independent travel: orientation (the ability to monitor one's position in relationship to the environment) and route guidance (the ability to determine a safe and appropriate route for reaching one's destination). As an orientation aid, the Global Positioning System (GPS) seems promising. For route guidance, a notebook computer or personal data assistant (PDA) equipped with speech input/output software, route planning software (artificial intelligence), and digital maps have been proposed (18,31,32). A voice operable, handheld GPS unit used in combination with obstacle detecting ETAs like the Laser Cane might constitute the ideal navigational aid for blind persons.

Several GPS equipped PDAs have recently become available. For example, the iQue 3600 ($600 from Garmin International Inc., Olathe, Kansas) is a handheld device that combines a PDA and mapping software with a built-in GPS receiver. The iQue 3600 uses the Palm operating system and offers a color screen and voice output turn-by-turn navigational guidance. For someone who already possesses a PDA (e.g., Palm Pilot or Microsoft's Pocket

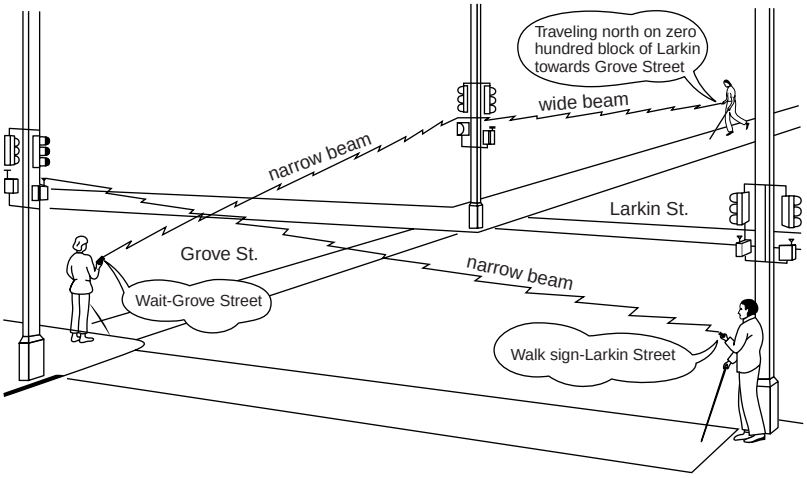


Figure 9. Talking Signs not only gives location information, but also tells the pedestrian the current status of the pedestrian cycle, aids in finding the cross-walk, and indicates the direction of the destination corner. (Reproduced from Ref. 29.)

PC), various third party software and GPS add-on units can be used.

While promising as a navigational aid for persons with severe visual impairments, GPS equipped portable PDAs (or notebook computers) have significant limitations. To fully appreciate these limitations, a brief review of how the Global Positioning System works (Fig. 10) would be apropos.

Global Positioning System (GPS) began some 30 years ago when Aerospace Corporation in Southern California studied ways to improve radio navigation systems for the military (33). Although GPS was not fully operational at the outbreak of the Persian Gulf War in January 1991, its exceptional performance in accurately locating fighting units evoked a strong demand from the military for its immediate completion.

Currently, 24 satellites of the GPS circle the earth every 12 h at a height of 20,200 km. Each satellite continuously transmits pseudorandom codes at 1575.42 and 1227.6 MHz. The orbital paths of the satellites and their altitude enable an unobstructed observer to see between five and eight satellites from any point on the earth. Signals from different visible satellites arrive at the GPS receiver with different time delays. The time delay needed to achieve coherence between the satellites' pseudorandom codes and the receiver's internally generated code equals the time-of-flight delay from a given satellite. GPS signals from at least four satellites are analyzed to determine the receiver's longitude, latitude, altitude (as measured from earth's center) and the user's clock error with respect to system time (33).

For civilian applications, position accuracy of a single channel receiver is about 100 m and its time accuracy is ~ 340 ns. Greater accuracy, usually within 1 m, can be achieved using differential GPS wherein signals from additional satellites are analyzed and/or the satellite signals are compared with and corrected by a GPS transceiver at a known fixed location (33).

At first glance, GPS signals seem fully able to meet the orientation needs of persons with severe visual impairments. The GPS signals are sufficiently accurate if combined with differential GPS and signals are immune to weather and are available at any time of the day, anywhere there's a line of sight to at least four GPS satellites. Lastly, a GPS receiver is relatively inexpensive, $< \$200$.

Unfortunately, just equipping blind persons with a voice-output GPS receiver for wayfinding outdoors is insufficient. The GPS signals are often unavailable or highly attenuated under bridges, inside natural canyons, and between tall buildings in urban areas. The altitude GPS information is generally not useful, and its longitudinal and latitudinal coordinates are useless when unaccompanied by local area maps (17). For college campuses or even major metropolitan areas, the location of major buildings and their entrances in terms of longitude and latitude coordinates are rarely available. Without these key pieces

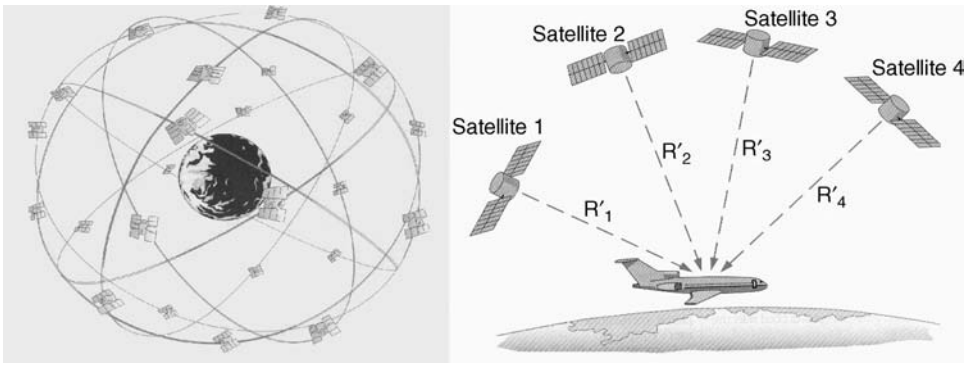


Figure 10. Synchronized signals from four satellites are analyzed by the mobile receiver to determine its precise position in three dimensions. The distances for the four satellites include an unknown error due to the inaccuracy of the receiver's clock and Doppler effects. (Reprinted from Ref. 33.)

of information, the GPS navigational aid is unable to offer directional guidance to the blind traveler.

INDOOR NAVIGATIONAL AIDS

One of the key characteristics of an ideal mobility aid is that the device reliably function indoors, outdoors, and within changing environments (Table 2). When used in combination with detailed local area maps, the GPS receiver and voice output could form the basis for a navigational aid. However, GPS signals may not be available at all times and are totally absent indoors. To function indoors, an electronic navigational aid will need to rely on some other set of electronic beacons as signposts.

For wayfinding within a large building, several investigators have borrowed the idea found in the Hansel and Gretel fairy tale about two children leaving a trail of bread crumbs to find their way home again. Instead of bread crumbs, Szeto (18) proposed placing small, low cost electronic beacons along corridors or at strategic locations (e.g., elevators, bathrooms, stairs) of buildings visited. Acting like personal pathmarkers, these radio frequency (RF) emitting electronic beacons would be detected by the associated navigational aid and guide the traveler back to a previous location or exit. To avoid confusion with other users, the electronic beacons could be keyed to work with one navigational aid.

Kulyukin et al. (34) recently studied the efficacy of using Radio Frequency Identification (RFID) tags in robot-assisted indoor navigational for the visually impaired. They described how strategically located, passive (nonpowered) RFID tags could be detected and identified by a RFID reader employing a square 200×200 mm antenna and linked to a laptop computer. In field tests, wall-mounted RFID tags responded to the spherical electromagnetic field from an RFID antenna at a distance of ~ 1.5 m. Since each tag is given a unique identifier, its location inside a building can be easily recalled and used to locate one's position inside a building.

In comparison to wall-mounted Talking Signs, the approach of Szeto (18) and Kulukin et al. (34) seems to be less costly and more flexible. Placing disposable electronic beacons in the hands of individual travelers does not require permanent retrofits of buildings, can be cost effective even for single users, and easily changes with the travel patterns of the user.

The electronic beacons and handheld electronic transceivers also should be economically feasible because they utilize a technology that's being developed for the mass market. World's largest retailer, Wal-Mart, has mandated 2008 as the year when all its suppliers must implement an RFID tracking system for their deliveries. It is likely that RFID tags, antenna, and handheld interrogators developed for inventory tracking can be adapted for use in an indoor navigational aid.

Although not yet a reality, a low cost, portable, handheld, indoor–outdoor mobility aid that embodies many of the features listed in Table 2 is clearly feasible. The needed technological infrastructure will soon be in place. For obstacle avoidance, the Laser Cane, Guide Cane, or their variants can be used. For indoor wayfinding and route guidance, the blind traveler could augment the cane or guide dog with a handheld voice output electronic navigational aid linked to strategically placed electronic beacons. For outdoor wayfinding, the blind traveler could augment the Laser Cane with a handheld mobility aid equipped with a GPS receiver, compass, local area maps, and wireless internet link.

The intelligent navigational aid just described would address the mobility needs of the blind by responding to voice commands; automatically detecting GPS signals or searching for the presence of electronic beacons; wirelessly linking to the local area network to obtain directory information; converting the GPS coordinates or the signals from electronic beacons into a specific location on a digital map; and, with the help of routing finding software, generating step-by-step directions to the desired destination.

FUTURE POSSIBILITIES

Of course, the ultimate assistive technology for overcoming the many problems of severe visual impairment would be an artificial eye. Since the mid-1990s, research by engineers, ophthalmologists, and biologists to develop a bionic eye have grown and artificial retina prototypes are nearing animal testing. An artificial eye would incorporate a small video camera to capture light from objects and transmit the image to a wallet-sized computer processor that in turn sends the image to an implant that would stimulate either the retina (35) or visual cortex (36).

Researchers at Stanford University recently announced progress toward an artificial vision system that can stimulate a retina with enough resolution to enable a visually impaired person to orient themselves toward objects, identify faces, watch television, read large fonts, and live independently (37). Their optoelectronic retinal prosthesis system is expected to stimulate the retina with resolution corresponding to a visual acuity of 20/80 by employing 2500 pixels per square millimeter. The researchers see the device as being particularly helpful for people left blind by retinal degeneration. Although such developments are exciting, tests with human subjects on practical but experimental prototypes won't likely occur for another 6–8 years (38).

What else the does the future hold in terms of assistive technology in general and mobility aids in particular? In an address to the CSUN 18th Annual Conference on Technology and Persons with Disabilities in 2003, futurist and U.S. National Medal of Technology recipient, Ray Kurzweil, presented his vision of the sweeping technological changes that he expected to take place over the next few decades (39). His comments are worthy of reflection and give cause for optimism.

With scientific and technological progress doubling every decade, Kurzweil envisions ubiquitous computers with always-on Internet connections, systems that would allow people to fully immerse themselves in virtual environments, and artificial intelligence embedded into Web sites by 2010. Kurzweil (39) expects the human brain to be fully reverse-engineered by 2020, which would result in computers with enough power to equal human intelligence. He forecasted the emergence of systems that provide subtitles for deaf people around the world, as well as listening systems geared

toward hearing-impaired users. Blind people would be able to take advantage of pocket-sized reading devices within a decade or have retinal implants that restore useful vision in 10–20 years. Kurzweil believed that people with spinal cord injuries would be able to resume fully functional lives by 2020, either through the development of exoskeletal robotic systems or techniques that bridged severed neural pathways, possibly by wirelessly transmitting nerve impulses to muscles. Even if one-half of what Kurzweil predicted became reality, the future of assistive technology for the blind is bright and an efficacious intelligent mobility aid for such persons will soon be commercially available.

BIBLIOGRAPHY

Cited References

1. Beck AF, Stern A, Uslan MM, Wiener WR, editors. Access to Mass Transit for Blind and Visually Impaired Travelers. New York: American Foundation for the Blind; 1990.
2. National Eye Institute and Prevent Blindness America®, Vision Problems in the U.S., 4th ed., 2002.
3. Arch Ophtha Imol. April 2004; 122.
4. Allen J. Electronic aids for the severely visually handicapped. CRC Crit Rev Bioeng 1971;1:137–167.
5. Servais SB. Visual Aids. In: Webster JG, Cook AM, Tompkins WJ, Vanderheiden GC, editors. Electronic Devices for Rehabilitation. New York: John Wiley & Sons, Medical; 1985. p 31–78.
6. Independent Living Aids, Inc. (No date) [Online] product catalog. Available at http://www.independentliving.com/home.asp. Accesse May 2005
7. Allen J. Linguistic-based algorithms offer practical text-to-speech systems. Speech Technol 1981;1(1):12–16.
8. Breen A. Speech synthesis models: a review. Elect Commun Eng J 1992;4(1):19–31.
9. O'Shaughnessy D. Interacting with computers by voice: Automatic speech recognition and synthesis. Proc IEEE 2003;91(9): 1272–1305.
10. Jackson G, et al. A single-chip text-to-speech synthesis device utilizing analog nonvolatile multi-level flash storage. IEEE J. Solid State Cir Nov 2002;37(11):1582–1592.
11. Thatcher J, et al. Constructing Accessible Websites, ISBN: 1904151000, New York: Glasshaus; 2002.
12. Matthews W. 13 rules for accessible web pages, August 07, 2000 of the Federal Computer Week. (No date). [Online]. Available at http:// www.fcw.com/fcw/articles/2000/0807/cov-access2-08-07-00.asp. Accessed March 2005.
13. Lazzaro JJ. Adaptive Technologies for Learning and Work Environments. 2nd ed., New York: The American Library Association; 2000.
14. Tucker A. Net surfing for those unable to see, Baltimore Sun, p. 1C. [Online] Available at http://www.baltimoresun.com/features/lifestyle/bal-to.blind16mar16,1,1345515.story?ctrack=1&cset=true. Accessed March 16, 2005.
15. Shao S. Mobility Aids For The Blind. In: Webster JG, Cook AM, Tompkins WJ, Vanderheiden JC, editors. Electronic Devices for Rehabilitation. New York: John Wiley & Sons, Medical; 1985. p. 79–100.
16. Farmer LW. Mobility Devices. In: Welsh RL, Blasch BB, editors. Foundation of Orientation and Mobility. New York: American Foundation for the Blind; 1980. p 206–209.
17. Bentzen BL. Orientation aids. In: Blasch B, Weiner W, Welsh W, editors. Foundations of Orientation and Mobility. 2nd ed. New York: American Foundation for the Blind; 1997. p 284–316.
18. Szeto AYJ. A navigational aid for persons with severe visual impairments: a project in progress. Proceeding of the 25th Annual International Conference IEEE Engineering and Medicine & Biology Society; Vol 25(2), Cancun, Mexico, Sep. 2003 p 1637–1639.
19. Nye PW, Bliss JC. Sensory aids for the blind: a challenging problem with lessons for the future. Proc IEEE 1970;58: 1878–1879.
20. Cook AM, Hssey SM. Assistive Technologies: Principles and Practice. 2nd ed., St. Louis, (MO): Mosby; 2002. p. 423–426.
21. Kay L. A sonar aid to enhance spatial perception of the blind: Engineering design and evaluation. Radio Elect Eng 1974;44(11):605–627.
22. Kuc R. Binaural sonar electronic travel aid provides vibrotactile cues for landmark, reflector motion and surface texture classification. IEEE Trans Biomed Eng Oct 2002; 49(10):1173–1180.
23. Shovel S, Ulrich I, Borenstein J. Computerized Obstacle Avoidance Systems for the Blind and Visually Impaired. In: Teodorescu HNL, Jain LC, editors. Intelligent Systems and Technologies in Rehabilitation Engineering. Boca Raton(FL): CRC Press; 2001.
24. Polaroid Corp, Ultrasonic Ranging System—Description, operation and use information for conducting tests and experiments with Polaroid's Ultrasonic Ranging System, Ultrasonic Components Group, 119 Windsor Street, Cambridge (MA).
25. National Safety Council, Pedestrian accidents, National Safety Council Accident Facts (Injury Statistics), 1998.
26. Szeto AYJ, Valerio N, Novak R. Audible pedestrian traffic signals: Part 1. Prevalence and impact. J Rehabilit R & D 1991;28(2):57–64.
27. Szeto AYJ, Valerio N, Novak R. Audible pedestrian traffic signals: Part 2. Analysis of sounds emitted. J Rehabilit R & D 1991;28(2):65–70.
28. Szeto AYJ, Valerio N, Novak R. Audible pedestrian traffic signals: Part 3. Detectability. J Rehabilit R & D 1991;28(2):71–78.
29. Farmer LW, Smith DL. Adaptive technology. In: Blasch B, Weiner W, Welsh R. Foundations of Orientation and Mobility. 2nd ed., New York: American Foundation for the Blind; 1997. p. 231–259.
30. Brabyn J, Crandall W, Gerrey W. Talking Signs®: A Remote Signage Solution for the Blind, visually Impaired and Reading Disabled, Proceeding of the 15th Annual International Conference in IEEE Engineering in Medicine & Biology Society; 1993; Vol. 15: p. 1309–1311.
31. Vogel S. A PDA-based navigational system for the blind. [Online], Available at http://www.cs.unc.edu/~vogel/IP/IP/IP_versions/IPfinal_SusanneVogel. Accessed Spring 2003.pdf.
32. Helal A, Moore SE, Ramachandran B. Drishti: An integrated navigation system for visually impaired and disabled. Procedings of the 5th International Symposiun on Wearable Computers, Zurich, Switzerland; October 2001; p. 149–155.
33. Getting IA. The Global Positioning System. IEEE Spectrum Dec. 1993;30(12):36–47.
34. Kulyukin V, Gharpure C, Nicholson J, Pavithran S. RFID in Robot-Assisted indoor Navigation for the Visually Impaired. Proceedings of the 2004 IEEE/RSJ International Conference on Intelligent Robots and Systems; Sept. 28–Oct. 2, 2004; Sendai, Japan, p 1979–1984.
35. Wyatt J, Rizzo J. Ocular implants for the blind. IEEE Spectrum May 1996;33(5):47–53.
36. Normann RA, Maynard EM, Guillory KS, Warren DJ. Cortical implants for the blind. IEEE Spectrum May 1996;33 (5):54–59.
37. Palanker D, Vankov A, Huie P, Baccus S. Design of a high-resolution optoelectronic retinal prosthesis. J Neural Eng 2005;2:105–120.

38. Braham R. Toward an artificial eye. IEEE Spectrum May 1996;33(5):20–21.
39. Kurzweil R. The future of intelligent technology and its impact on disabilities. J Visual Impairment Blindness Oct 2003;97(10):582–585.

Reading List

Cook AM, Hussey SM. Assistive Technologies: Principles and Practice. 2nd ed., St. Louis (MO): Mosby, Inc.; 2002. A thorough text on assistive technologies that is especially suited for the rehabilitation practitioner or those in allied health.

Smith RV, Leslie JH Jr., editors. Rehabilitation Engineering. Boca Rotan (FL): CRC Press; 1990. Contains diverse articles that should be of particular interest to practitioners in the rehabilitation field although several of the articles present definitive state-of-the-art information on rehabilitation engineering.

Webster JG, Cook AM, Tompkins WJ, Vanderheiden GC, editors. Electronic Devices for Rehabilitation. New York: John Wiley & Sons Inc. Medical; 1985. Though somewhat dated, this book offers a comprehensive overview of rehabilitation engineering and describes many of the design issues that underlie various types of assistive devices. A useful introductory text for undergraduate engineering students interested in rehabilitation.

Golledge RG. Wayfinding Behavior: Cognitive Mapping and Other Spatial Processes. Baltimore: John Hopkins University Press; 1999. A good reference that covers the cognitive issues of wayfinding behavior in blind and sighted humans.

Teodorescu HNL, Jain LC, editors. Intelligent Systems and Technologies in Rehabilitation Engineering. Boca Rotan (FL): CRC Press; 2000. A compendium of technical review articles covering intelligent technologies applied to retinal prosthesis, auditory & cochlear prostheses, upper and lower limb orthoses/prostheses, neural prostheses, pacemakers, and robotics for rehabilitation.

Yonaitis RB. Understanding Accessibility-A Guide to Achieving Compliance on Websites and Intranets, ISBN: 1-930616-03-1, HiSoftware, 2002. This is a free booklet in electronic form for complying with the Federal government's "Section 508" of the Workplace Rehabilitation Act (amendments of 1998). The book gives a brief and clear discussion of accessibility testing and how to integrate this activity into web design and related tasks.

Blasch B, Weiner W, Welsh R. Foundations of Orientation and Mobility. 2nd ed., New York: American Foundation for the Blind; 1997. A useful book for general background regarding the issues of orientation and mobility.

IEEE Spectrum, Vol. 33(5), May 1996, carries six special reports on the development of an artificial eye. Articles in this issue examine physiology of the retina, neural network signal processing, electrode array design, and sensor technology.

Journal of Visual Impairment and Blindness, Vol. 97(10), Oct. 2003, is a special issue that focused on the impact of technology on blindness.

Speech Technology, a magazine published bimonthly by AmCom Publications, 2628 Wilhite Court, Suite 100, Lexington, KY 450503. This magazine regularly covers the development and implementation of technologies that underlie speech recognition and speech generation. For example, its March/April 2005 issue contained articles on the following topics: guide to speech standards; applications of transcription; role of speech in healthcare, embedding speech into mobile devices, technology trends, new products, and speech recognition software.

See also COMMUNICATION DEVICES; ENVIRONMENTAL CONTROL; MOBILITY AIDS; VISUAL PROSTHESES.

BLOOD BANKING. See BLOOD COLLECTION AND PROCESSING.

BLOOD CELL COUNTERS. See CELL COUNTERS, BLOOD.

BLOOD COLLECTION AND PROCESSING

TERESA M. TERRY
JOSEPHINE H. COX
Walter Reed Army Institute of Research
Rockville, Maryland

INTRODUCTION

Phlebotomy may date back to the Stone Age when crude tools were used to puncture vessels to allow excess blood to drain out of the body (1). This purging of blood, subsequently known as blood letting, was used for therapeutic rather than diagnostic purposes and was practiced through to modern times. Phlebotomy started to be practiced in a more regulated and dependable fashion after the Keidel vacuum tube for the collection of blood was manufactured by Hynson, Wescott, and Dunning. The system consisted of a sealed ampoule with or without culture medium connected to a short rubber tube with a needle at the end. After insertion onto the vein, the stem of the ampoule was crushed and the blood entered the ampoule by vacuum. Although effective, the system did not become popular until evacuated blood collection systems started to be used in the mid-twentieth century. With evacuated blood collection systems came a new interest in phlebotomy and blood drawing techniques and systems. A lot of technical improvements have been made, not only are needles smaller, sharper, and sterile, they are also less painful. The improved techniques of obtaining blood samples assure more accurate diagnostic results and less permanent damage to the patient. Today, the main purpose of phlebotomy synonymous with venipuncture is to obtain blood for diagnostic testing.

Venipuncture Standards and Recent Standard Changes

The Clinical and Laboratory Standards Institute (CLSI, formerly the National Committee for Clinical Laboratory Standards, NCCLS) develops guidelines and sets standards for all areas of the laboratory (www.CLSI.org). Phlebotomy program approval as well as certification examination questions are based on these important national standards. Another agency that affects the standards of phlebotomy is the College of American Pathologists (CAP; www.CAP.org). This national organization is an outgrowth of the American Society of Clinical Pathologists (ASCP). The membership in this specialty organization is made up of board-certified pathologists only and offers, among other services, a continuous form of laboratory inspection by pathologists. The CAP Inspection and Accreditation Program do not compete with the Joint Commission on Accreditation of Health Care Organizations

(JCAHO) accreditation for health care facilities, because it was designed for pathology services only.

The CLSI have published the most current research and industry regulations on standards and guidelines for clinical laboratory procedures (2,3). The most significant changes to specimen collection are (*1*) collectors are now advised to discard the collection device without disassembling it, this reflects the Occupational Safety and Health Administration's (OSHA) mandate against removing needles from tube holders; (*2*) the standard now permits gloves to be applied just prior to site preparation instead of prior to surveying the veins; (*3*) collectors are advised to inquire if the patient has a latex sensitivity; (*4*) sharp containers should be easily accessible and positioned at the point of use; (*5*) there is a caution recommended against the use of ammonia inhalants on fainting patients in case the patient is asthmatic; (*6*) collectors must attempt to locate the median cubital vein on either arm before considering alternative veins due to the proximity of the basilica vein to the brachial artery and the median nerve; (*7*) forbids lateral needle relocation in an effort to access the basilica vein to avoid perforating or lacerating the brachial artery; (*8*) immediate release of tourniquet "if possible" upon venous access to prevent the effects of hemoconcentration from altering test results.

The Role of the Phlebotomist Today

Professionalism. Phlebotomists are healthcare workers and must practice professionalism and abide by state and federal requirements. A number of agencies have evolved offering the phlebotomist options for professional recognition (1). Certification is a process that indicates the completion of defined academic and training requirements and the attainment of a satisfactory score on a national examination. Agencies that certify phlebotomists and the title each awards include the following: American Society of Clinical Pathologists (ASCP): Phlebotomy Technician, PBT (ASCP); American Society for Phlebotomy Technology (ASPT): Certified Phlebotomy Technician, CPT (ASPT); National Certification Agency for Medical Laboratory Personnel (NCA): Clinical Laboratory Phlebotomist (CLP) (NCA); National Phlebotomy Association (NPA); Certified Phlebotomy Technician, CPT (NPA). Licensure is defined as a process similar to certification, but at the state or local level. A license to practice a specific trade is granted through examination to a person who can meet the requirements for education and experience in that field. Accreditation and approval of healthcare training programs provides an individual with an indication of the quality of the program or institution. The accreditation process involves external peer review of the educational program, including an on-site survey to determine if the program meets certain established qualification or educational standards referred to as 'essentials'. The approval process is similar to accreditation; however, programs must meet educational—standards and competencies—rather than essentials, and an on-site survey is not required.

Public Relations and Legal Considerations. The Patient's Bill of Rights was originally published in 1975 by the American Hospital Association. The document, while not legally binding, is an accepted statement of principles that guides healthcare workers in their dealings with patients. It states that all healthcare professionals, including phlebotomists, have a primary responsibility for quality patient care, while at the same time maintaining the patient's personal rights and dignity. Two rights especially pertinent to the phlebotomist are the right of the patient to refuse to have blood drawn and the right to have results of lab work remain confidential. Right of Privacy: "An individual's right to be let alone, recognized in all United States jurisdictions, includes the right to be free of intrusion upon physical and mental solitude or seclusion and the right to be free of public disclosure of private facts. Every healthcare institution and worker has a duty to respect a patient's or client's right of privacy, which includes the privacy and confidentiality of information obtained from the patient–client for purposes of diagnosis, medical records, and public health reporting requirements. If a healthcare worker conducts tests on or publishes information about a patient–client without that person's consent, the healthcare worker could be sued for wrongful invasion of privacy, defamation, or a variety of other actionable torts." In 1996, the Health Insurance Portability and Accountability Act (HIPAA) law was signed. It is a set of rules to be followed by health plans, doctors, hospitals, and other healthcare providers. Patients must be able to access their record and correct errors and must be informed of how their personal information will be used. Other provisions involve confidentiality of patient information and documentation of privacy procedures.

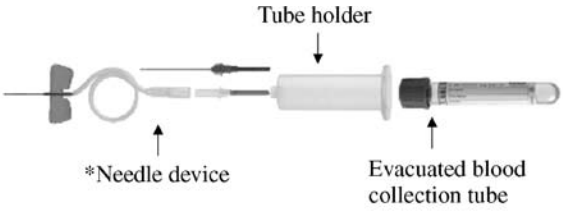
SAFETY

Universal Precautions

An approach to infection control that is mandated by federal and state laws is the so-called Universal Precautions. The guidelines for Universal Precautions are outlined by OSHA (www.OSHA.gov). According to the concept of Universal Precautions, all human blood and certain human body fluids are treated as if known to be infectious for human immunodeficiency virus (HIV), hepatitis B virus (HBV), and other blood borne pathogens. For blood collections, the use of needles with a safety device or a needle integrated into a safety device and the use of gloves is now mandatory in most institutions. Biohazard material should be disposed of in an appropriately labeled biohazard container. Needles and other sharp instruments should be disposed of in rigid puncture-resistant biohazard containers.

First Aid Procedures

Most phlebotomy programs require cardio pulmonary resuscitation (CPR) certification as a prerequisite or include it as part of the course and in the event of an emergency situation: basic First Aid Procedures should be performed by the phlebotomist. These procedures are not in the scope of this article and training needs to be performed by qualified experts.



*A butterfly with luer adapter is shown as well as a standard needle

Figure 1. Basic components of the Evacuated Blood Collection System.

BLOOD COLLECTION SYSTEM AND EQUIPMENT

Blood Collection System

The components of the Evacuated Blood Collection System are shown in Fig. 1. The system consists of the following;

Plastic evacuated collection tube: The tubes are designed to fill with a predetermined volume of blood by vacuum. The rubber stoppers are color coded according to the additive that the tube contains (see Table 1). Evacuated collection devices are supplied by many vendors worldwide. These evacuated collection devices use similar color coding systems, proprietary additives, and recommended uses. Various sizes are available.

Tube holder (single use): For use with the evacuated collection system.

Needles (also available with safety device): The gauge number indicates the bore size: the larger the gauge number, the smaller the needle bore. Needles are available for evacuated systems and for use with a syringe, single draw, or butterfly system.

Additional Materials

Tourniquet: Wipe off with alcohol and replace frequently. Nonlatex tourniquets are recommended.

Table 1. Tube Guide[a]

Tube Top Color	Additive	Inversions at Blood Collection[a]	Laboratory Use
Gold or Red/Black	Clot activator Gel for serum separation	5	Tube for serum determinations in chemistry. Blood clotting time: 30 min
Light Green or Green/Gray	Lithium heparin Gel for plasma separation	8	Tube for plasma determinations in chemistry
Red	Clot activator	5	Tube for serum determination in chemistry, serology, and immunohematology testing
Orange or Gray/Yellow	Thrombin	8	Tube for stat serum determinations in chemistry. Blood clotting occurs in < 5 min
Royal Blue	Clot activator K_2EDTA, where EDTA = ethylenediaminetetraacetic acid	5 8	Tube for trace-element, toxicology and nutritional chemistry determinations.
Green	Sodium heparin Lithium heparin	8 8	Tube for plasma determination in chemistry
Gray	Potassium oxalate/sodium fluoride Sodium fluoride/Na_2EDTA Sodium fluoride (serum tube)	8 8 8	Tube for glucose determination. Oxalate and EDTA anticoagulants will give plasma samples. Sodium fluoride is the antiglycolytic agent
Tan	K_2EDTA	8	Tube for lead determination. This tube is certified to contain $< 0.01\ \mu g \cdot mL^{-1}$ lead
Lavender	Spray-coated K_2EDTA	8	Tube for whole blood hematology determination and immunohematology testing
White	K_2EDTA with gel	8	Tube for molecular diagnostic test methods such as polymerase chain reaction (PCR) and/or DNA amplification techniques.
Pink	Spray-coated K_2EDTA	8	Tube for whole blood hematology determination and immunohematology test. Designed with special cross-match label for required patient information by the AABB[b]
Light Blue	Buffered sodium citrate (3.2%) Citrate, theophylline, adenosine, dipyridamole (CTAD)	3	Tube for coagulation determinations. The CTAD for selected platelet function assays and routine coagulation determination

[a]Reproduced from Becton Dickinson www.bd.com/vacutainer. Evacuated collection devices made by other manufacturers use similar color coding systems and additives. Recommended inversion times and directions for use are provided by each supplier.

[b]AABB = American Association of Blood Banks.

Gloves: Worn to protect the patient and the phlebotomist. Nonlatex gloves are recommended.

Antiseptics–Disinfectants: 70% isopropyl alcohol or iodine wipes (used if blood culture is to be drawn).

Sterile gauze pads: For application on the site from which the needle is withdrawn.

Bandages: Protects the venipuncture site after collection.

Disposal containers: Needles should never be broken, bent, or recapped. Needles should be placed in a proper disposal unit immediately after use.

Syringe: May be used in place of evacuated collection system in special circumstances.

Permanent marker or pen: To put phlebotomist initials, time, and date of collection on tube as well as any patient identification information not provided by test order label.

BEST SITES FOR VENIPUNCTURE

The most common sites for venipuncture are located in the antecubital (inside elbow) area of the arm (see Fig. 2). The primary vein used is the median cubical vein. The basilica and cephalic veins can be used as a second choice. Although the larger and fuller median cubital and cephalic veins of the arm are used most frequently, wrist and hand veins are also acceptable for venipuncture. Certain areas are to be avoided when choosing site such as; (*1*) Skin areas with extensive scars from burns and surgery (it is difficult to puncture the scar tissue and obtain a specimen); (*2*) the upper extremity on the side of a previous mastectomy (test results may be affected because of lymphedema); (*3*) site of a hematoma (may cause erroneous test results). If another site is not available, collect the specimen distal to the hematoma; (*4*) Intravenous therapy (IV)/blood transfusions (fluid may dilute the specimen, so collect from the opposite arm if possible); (*5*) cannula/fistula/heparin lock (hospitals have special safety and handling policies regarding these devices). In general, blood should not be drawn from an arm with a fistula or cannula without consulting the attending physician; (*6*) edematous extremities (tissue fluid accumulation alters test results).

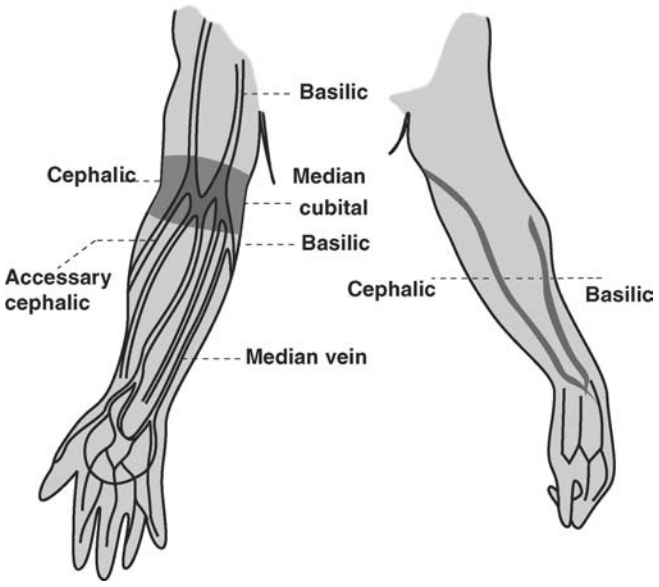


Figure 2. Venipuncture sites.

ROUTINE PHLEBOTOMY PROCEDURE

Venipuncture is often referred to as "drawing blood". Most tests require collection of a blood specimen by venipuncture and a routine venipuncture involves the following steps *Note*: The following steps were written using guidelines established by the CLSI/NCCLS (3).

1. **Prepare Order.** The test collection process begins when the physician orders or requests a test to be performed on a patient. All laboratory testing must be requested by a physician and results reported to a physician. The form on which the test is ordered and sent to the lab is called the test requisition. The requisition may be a computer-generated form or a manual form.
2. **Greet and Identify Patient.** Approach the patient in a friendly, calm manner. Identify yourself to the patient by stating you name. Provide for their comfort as much as possible. The most important step in specimen collection is patient identification. When identifying a patient, ask the patient to state their name and date of birth. Outpatients can use an identification card as verification of identity. Even if the patient has been properly identified by the receptionist, the phlebotomist must verify the patient's ID once the patient is actually called into the blood drawing area. The phlebotomist should ask for two identifiers that match the test requisition form (e.g., name and social security or name and date of birth).
3. **Verify Diet Restrictions and Latex Sensitivity.** Once a patient has been identified, the phlebotomist should verify that the patient has followed any special diet instructions or restrictions. The phlebotomist should also inquire about the patients' sensitivity to latex.

   Assemble Supplies: See the section on Blood Collection System and Equipment

   Position Patient

   A patient should be either seated or lying down while having blood drawn. The patient's arm that will be used for the venipuncture should be supported firmly and extended downward in a straight line.

4. **Apply Tourniquet.** A tourniquet is applied to increase pressure in the veins and aid in vein selection. The tourniquet is applied 3–4 in. (7.62–10.18 cm) above the intended venipuncture site. Never leave the tourniquet in place longer than 1 min.

5. **Select a Vein.** Palpate and trace the path of veins in the antecubital (inside elbow) area of the arm with the index finger. Having a patient make a fist will help make the veins more prominent. Palpating will help to determine the size, depth, and direction of the vein. The main veins in the antecubital area are the median cubical, basilica, and cephalic (see the section; Best Sites for Venipucture). Select a vein that is large and well anchored.
6. **Put on Gloves.** Properly wash hands followed by glove application.
7. **Cleanse Venipuncture Site.** Clean the site using a circular motion, starting at the center of the site and moving outward in widening concentric circles. Allow the area to air dry.
8. **Perform Venipuncture.** Grasp patients arm firmly to anchor the vein. Line the needle up with the vein. The needle should be inserted at a 15–30° angle BEVEL UP. When the needle enters the vein, a slight "give" or decrease in resistance should be felt. At this point, using a vacuum tube, slightly, with firm pressure, push the tube into the needle holder. Allow tube to fill until the vacuum is exhausted and blood ceases to flow to assure proper ratio of additive to blood. Remove the tube, using a twisting and pulling motion while bracing the holder with the thumb. If the tube contains an additive, mix it immediately by inverting it 5–10 times before putting it down.
9. **Order of Draw.** Blood tubes are drawn in a particular order to ensure integrity of each sample by lessening the chances of anticoagulants interference and mixing. The order of draw also provide a standardized method for all laboratories (3,4).

   Blood Cultures: With sodium polyanethol sulfonate anticoagulant and other supplements for bacterial growth.
   Light Blue: Citrate Tube (*Note*: When a citrate tube is the first specimen tube to be drawn, a discard tube should be drawn first). The discard tube should be a nonadditive or coagulation tube.
   Gold or Red/Black: Gel Serum Separator Tube, no additive.
   Red: Serum Tube, no additive.
   Green: Heparin Tube.
   Light Green or Green/Gray: Gel Plasma Separator Tube with Heparin.
   Lavender: EDTA Tube.
   Gray: Fluoride (glucose) Tube.

10. **Release the Tourniquet.** Once blood begins to flow the tourniquet may be released to prevent hemoconcentration.
11. **Place the Gauze Pad.** Fold clean gauze square in half or in fourths and place it directly over the needle without pressing down. Withdraw the needle in one smooth motion, and immediately apply pressure to the site with a gauze pad for 3–5 min, or until the bleeding has stopped. Failure to apply pressure will result in leakage of blood and hematoma formation. Do not bend the arm up, keep it extended or raised.
12. **Remove and Dispose of the Needle.** Needle should be disposed of immediately by placing it and the tube holder in the proper biohazard sharps container. Dispose of all other contaminated materials in proper biohazard containers.
13. **Bandage the Arm.** Examine the patients arm to assure that bleeding has stopped. If bleeding has stopped, apply an adhesive bandage over the site.
14. **Label Blood Collection Tubes.** Specimen tube labels should contain the following information: patient's full name, patient's ID number, date, time, and initials of the phlebotomist must be on each label of each tube.
15. **Send Blood Collection Tube to be Processed.** Specimens should be transported to the laboratory processing department in a timely fashion. Some tests may be compromised if blood cells are not separated from serum or plasma within a limited time.

SPECIMEN PROCESSING

Processing of blood is required in order to separate out the components for screening, diagnostic testing, or for therapeutic use. This section will concentrate primarily on processing of blood for screening purposes and diagnostic testing. An overview of the main blood processing procedures, specimen storage, and common uses for each of the components is provided in Table 2. Because there are many different blood components and many different end uses for these components, the list is not comprehensive and the reader should refer to other specialized literature for further details. The OSHA regulations require laboratory technicians to wear protective equipment (e.g., gloves, labcoat, and protective face gear) when processing specimens. Many laboratories mandate that such procedures are carried out in biosafety cabinets.

Whole Blood Processing

Because whole blood contains all but the active clotting components, it has the ability to rapidly deteriorate and all blood components are subject to chemical, biological, and physical changes. For this reason, whole blood has to be carefully handled and any testing using whole blood has to be performed as soon as possible after collection to ensure maximum stability. Whole blood is typically used for the complete blood count (CBC). The test is used as a broad screening test to check for such disorders as anemia, infection, and many other diseases (www.labtestsonline.org). The CBC panel typically includes measurement of the following: white blood, platelet and red blood cell count, white blood cell differential and evaluation of the red cell compartment by analysis of hemoglobin and hematocrit, red cell distribution width and mean corpuscular volume, and mean corpuscular hemoglobin. The CBC assays are now routinely performed with automated

Table 2. Blood Processing Procedures and Specimen Storage

Component	Processing	Short-Term Storage	Long-Term Storage	Uses
Red blood cells	Gravity and/or centrifigation	~1 month at 4 °C	Frozen up to 10 years	Transfusion
Plasma	Gravity and/or centrifigation	Use immediately	Frozen up to 7 years	Serology, diagnostics, immune monitoring source of biologics
Serum	Clotting and centrifugation	Use immediately	Frozen up to 7 years	Serology, diagnostics, immune monitoring source of biologics
Platelets	Plasma is centrifuged to enrich for platelet fraction	Five days at room temperature	Cannot be cryopreserved	Transfusion
Granulocytes	Centrifigation and separation from red blood cells	Use within 24 h	Cryopreserved in liquid nitrogen[a]	Transfusion
Peripheral blood mononuclear cells	Ficoll–hypaque separation	Use immediately	Cryopreserved in liquid nitrogen[a]	Immune monitoring, specialized expansion and reinfuison
Albumin, immune globulin, specific immune globulins, and clotting factor concentrates	Specialized processing, fractionation and separation	Not applicable	Variable	Multiple therapeutic uses

[a]Although it has been shown that cells can be stored indefinitely in liquid nitrogen, the functionality of the cells would have to be assessed and storage lengths determined for each type of use proposed.

analyzers in which capped evacuated collection devices are mixed and pierced through the rubber cap. Whole blood drawn in EDTA (lavender) tubes are usually used, although citrate (blue top) vacutainers will also work (although the result must be corrected because of dilution). Blood is sampled and diluted, and moves through a tube thin enough that cells pass by one at a time. Characteristics about the cell are measured using lasers or electrical impedance. The blood is separated into a number of different channels for different tests.

The CBC technology has expanded in scope to encompass a whole new field of diagnostics, namely, analytical cytometry. Analytical cytometry is a laser-based technology that permits rapid and precise multiparameter analysis of individual cells and particles from within a heterogeneous population of blood or tissues. Analytical cytometry is now routinely used for diagnosis of different pathological states. This technique can be used to examine cell deoxyribonucleic acid (DNA) and ribonucleic acid (RNA) content, cell-cycle distribution, cellular apoptosis, tumor ploidy, cell function measurements (i.e., oxidative metabolism, phagocytosis), cellular biochemistry (i.e., intracellular pH, calcium mobilization, membrane potential, microviscosity, glutathione content), and fluorescence image analysis of individual blood cells. Since the blood is truly a window on what happens in the body, it is possible to use blood samples for a wide array of diagnostic and research purposes.

A second important use for whole blood is in the setting of HIV infection and treatment as a way to monitor CD4 T cell counts and percentages. These single or multiplatform tests use fresh whole blood with EDTA as anticoagulant (< 18 h after collection) samples are run with or without lysis of red cells and fixation of the lymphocytes. There are several different companies that make specialized equipment for enumeration of CD4 cell counts, the basic principle of which is to use a fluorescently tagged anti CD4 cell surface marker. The results for the CD4 or other subset are expressed as a percentage of total gated lymphocytes. In order to determine the absolute CD4 cell counts, the percent CD4 must be multiplied by an absolute lymphocyte count derived from a hematology analyzer or by an integrated volumetric analysis method (5–7).

Another common use for whole blood is for the detection of secretion of cytokines from antigen or mitogen stimulated lymphocytes. This assay can provide information on a patients T cell response to pathogens [e.g., cytomegalo virus (CMV) and Epstein Ban virus (EBV), HIV, and tuberculosis (TB)]. The technique can also be used to monitor vaccine induced responses or responses to immunotherapy. Whole blood is drawn into heparin, 0.5–1 mL of blood is stimulated with antigen of interest and costimulatory antibodies in the presence of Brefeldin-A. The latter inhibits transport of proteins from the Golgi so that secreted cytokines accumulate inside the cell. The samples are incubated at 37 °C for 6 h, after which they can be placed at 4 °C overnight or processed immediately. The samples are treated with EDTA to reduce clumping, red cells are lysed, and the sample is fixed by addition of paraformaldehyde. At this stage, the samples can be stored frozen for up to 4 months prior to detection of cell surface markers and intracellular cytokines by flow cytometry (8).

Serum Processing

Because of the ease of performing serum separation and the fact that so many tests rely on the use of serum, the technique has become routine in clinical and diagnostic laboratories. Specimens are drawn into tubes that contain no additives or anticoagulants (Table 1). Two commonly used tubes are the red serum collection tubes or commercially available serum separation tubes. Serum is obtained

by drawing the blood into a red top or the serum separator tube, allowing it to clot, and centrifuging to separate the serum. The time allowed for clotting depends on the ambient temperature and the patient sample. The typical recommendation is to allow the tube to clot for 20–30 min in a vertical position. A maximum of 1 h should suffice for all samples except those from patients with clotting disorders. Once the clot has formed, the sample is centrifuged for a recommended time of 10 min at 3000 revolutions per minute (rpm). The serum is transferred into a plastic transport tube or for storage purposes into a cryovial. Many tests collected in the serum separator tubes do not require transferring the supernatant serum unless the serum is to be stored frozen. Specimens transported by mail or stored >4 h should be separated from the clot and placed into a transport tube. Polypropylene plastic test tubes or cryovials are more resistant to breakage than most glass or plastic containers, especially when specimens are frozen. Caution needs to be observed with serum separator tubes for some tests since the analyte of interest may absorb to the gel barrier. Erroneous results may be obtained if the serum or plasma is hemolyzed, lipemic, or icteric. As eloquently described by Terry Kotrla, phlebotomist at Austin Community College these conditions cause specimen problems. (www.austin.cc.tx.us/kotrla/ PHBLab15SpecimenProcessingSum03.pdf).

1. **Hemolysis** is a red or reddish color in the serum or plasma that will appear as a result of red blood cells rupturing and releasing the hemoglobin molecules. Hemolysis is usually due to a traumatic venipuncture (i.e., vein collapses due to excessive pressure exerted with a syringe, " digging" for veins, or negative pressure damages innately fragile cells. Gross hemolysis (serum or plasma is bright red) affects most lab tests performed and the specimen should be recollected. Slight hemolysis (serum or plasma is lightly red) affects some tests, especially serum potassium and LDH (lactate dehydrogenase). Red blood cells contain large amounts of both of these substances and hemolysis will falsely elevate their measurements to a great extent. In addition to hemolysis caused during blood draw procedures, blood collection tubes (for serum and or whole blood) that are not transported correctly or in a timely fashion to the processing laboratory may be subject to hemolysis. Extremes of heat and cold in particular can cause red blood cells to lyze and sheering stresses caused by shaking of the specimens during transport may cause lysis. Finally, incorrect centrifugation temperatures and speeds may cause hemolysis of red blood cells.
2. **Icterus**. Serum or plasma can be bright yellow or even brownish due to either liver disease or damage or excessive red cell breakdown inside the body. Icterus can, like hemolysis, affect many lab tests, but unfortunately, recollection is not an option since the coloration of the serum or plasma is due to the patient's disease state.
3. **Lipemia**. Occasionally, serum or plasma may appear milky. Slight milkiness may be caused when the specimen is drawn from a nonfasting patient who has eaten a heavy meal. A thick milky appearance occurs in rare cases of hereditary lipemia.

Both for serum and plasma there are documented guidelines for specimen handling dependent on which analyte, is being examined. The kinds of tests that can be done on blood samples is ever expanding and includes allergy evaluations, cytogenetics, cytopathology, histopathology, molecular diagnostics, tests for analytes, viruses, bacteria, parasites, and fungi. Incorrect preparation, shipment, and storage of specimens may lead to erroneous results. The guidelines for preparing samples can be obtained from the CLSI (9). Diagnostic testing laboratories (e.g., Quest diagnostics) provide comprehensive lists of the preferred specimen type, transport temperature, and rejection criteria (www.questest.com).

Plasma Processing

Specimens are drawn into tubes that contain anticoagulant (Table 1.). The plasma is obtained by drawing a whole blood specimen with subsequent centrifugation to separate the plasma. Plasma can be obtained from standard blood tubes containing the appropriate anticoagulant or from commercially available plasma separation tubes. The plasma separation tubes combine spray-dried anticoagulants and a polyester material that separates most of the erythrocytes and granulocytes, and some of the lymphocytes and monocytes away from the supernatant. The result is a convenient, safe, single-tube system for the collection of whole blood and the separation of plasma. Samples can be collected, processed, and transported *in situ* thereby reducing the possibility of exposure to bloodborne pathogens at the collection and sample processing sites. One drawback is that plasma prepared in a plasma separation tube may contain a higher concentration of platelets than that found in whole blood. For plasma processing, after drawing the blood, the tube for plasma separation must be inverted five to six times to ensure adequate mixing and prevent coagulation. The recommended centrifugation time is at least 10 min at 3000 rpm. Depending on the tests required, plasma specimens may be used immediately, shipped at ambient or cooled temperatures, or may require freezing. The plasma is transferred into a plastic transport tube or for storage purposes into a cryovial. Some tests require platelet poor plasma, in which case the plasma is centrifuged at least two times.

Processing and Collection of Peripheral Blood Mononuclear Cells (PBMC) from Whole Blood

Peripheral blood mononuclear cells are a convenient source of white blood cells, T cells comprise $\sim 70\%$ of the white cell compartment and are the work-horses of the immune system. These T cells play a crucial role in protection from or amelioration of many human diseases and can keep tumors in check. The most readily accessible source of T

cells is the peripheral blood. Thus collection, processing, cryopreservation, storage, and manipulation of human PBMC are all key steps for assessment of vaccine and disease induced immune responses. The assessment of T cell function in assays may be affected by procedures beginning with the blood draw through cell separation, cryopreservation, storage, and thawing of the cells prior to the assays. Additionally, the time of blood collection to actual processing for lymphocyte separation is critical. Procedures for PBMC collection and separation are shown in Table 3 along with potential advantages and disadvantages.

When conducting cellular immunology assays, the integrity of the PBMC, especially the cellular membranes, is critical for success. A correct cellular separation process yields a pure, highly viable population of mononuclear cells consisting of lymphocytes and monocytes, minimal red blood cell and platelet contamination, and optimum functional capacity. The standard method for separation of PBMC is the use of Ficoll-hypaque gradients as originally described by Boyum in 1968 (10). A high degree of technical expertise is required to execute the procedure from accurate centrifuge rpm and careful removal of the cellular interface to avoid red cell contamination. Within the last 10 years, simplified separation ficoll procedures have largely replaced the standard ficoll method, two such procedures are outlined below (Adapted from Ref. 11) and in Fig. 3. The simplicity of these methods, superior technical reliability, reduced interperson variability, faster turnaround, and higher cell yields makes these the methods of choice.

The Cell Preparation Tube (CPT) method is described below and in greater detail in literature provided by Becton Dickinson (http://www.bd.com/vacutainer/products/molecular/citrate/procedures.asp). Vacutainer cell preparation tubes (VACUTAINER CPT tubes, Becton Dickinson) provide a convenient, single-tube system for the collection and separation of mononuclear cells from whole blood. The CPT tube is convenient to use and results in high viability of the cells after transportation. The blood specimens in the tubes can be transported at ambient temperature, as the gel forms a stable barrier between the anticoagulated blood and ficoll after a single centrifuge step. Cell separation is performed at the processing–storage laboratory using a single centrifugation step. This reduces the risk of sample contamination and eliminates the need for additional tubes and processing reagents. In many instances, and in particular when biosafety level 2 (BL2) cabinets are not available on site, the CPT method is useful because the centrifugation step can be done on site and the remaining processing steps can be performed after shipment to a central laboratory within the shortest time possible, optimally within 8 h. The central laboratory can complete cell processing in a BL2 cabinet and set up functional assays or cryopreserve the samples as needed.

Centrifuge speed is critical for PBMC processing. The centrifugal force is dependent on the radius of the rotation of the rotor, the speed at which it rotates, and the design of the rotor itself. Centrifugation procedures are given as xg measures, since rpm and other parameters will vary with the particular instrument and rotor used. The rpms may be calculated using the following formula where $r =$ radius of rotor $g =$ gravity; $g = 1.12\,r\,(\text{rpm}/1000)^2$. This conversion can be read-off a nomogram chart available readily online or in centrifuge maintenance manuals. Typically laboratory centrifuges can be programmed to provide the correct rpm.

Protocol 1. Separation of PBMC Using CPT Tubes

1. **Materials and Reagents:** Vacutainer CPT tubes (Becton Dickinson); Sterile Phosphate Buffered Saline (PBS) without Ca^+ and Mg^+, supplemented with antibiotics (Penicillin and Streptomycin); Sterile RPMI media containing 2% fetal bovine serum (FBS) and supplemented with antibiotics.

   The CPT tubes are sensitive to excessive temperature fluctuations, resulting in deterioration of the gel and impacting successful cellular separation. This problem is particularly serious in tropical countries where ambient storage temperatures may be $> 25\,^{\circ}\text{C}$. Following PBMC separation, one

Table 3. Stages and Variables in the Separation of PBMC from Whole Blood

Procedure/Technology	Alternatives	Advantages	Disadvantages
PBMC collection	Heparin	Greater cellular stability than EDTA	Impacts DNA isolation. Plasma from whole blood cannot be used for PCR based assays[a]
	EDTA		Time dependent negative impact on T cell responses
	Sodium Citrate	Greater cellular stability than EDTA	
PBMC separation	Standard Ficoll		Technically challenging Time consuming
	CPT	Rapid Technically easy and less inter-person variability Blood is drawn into same tube that is used for separation	Subject to temperature fluctuations manifested by gel deterioration and contamination in PBMC fraction.
	Accuspin/Leucosep	Rapid Technically easy and less inter-person variability	

[a]The inhibitory effects of heparin on DNA isolation can be removed by incubation of plasma or other specimens with silicon glass beads or by heparinase treatment prior to DNA extraction.

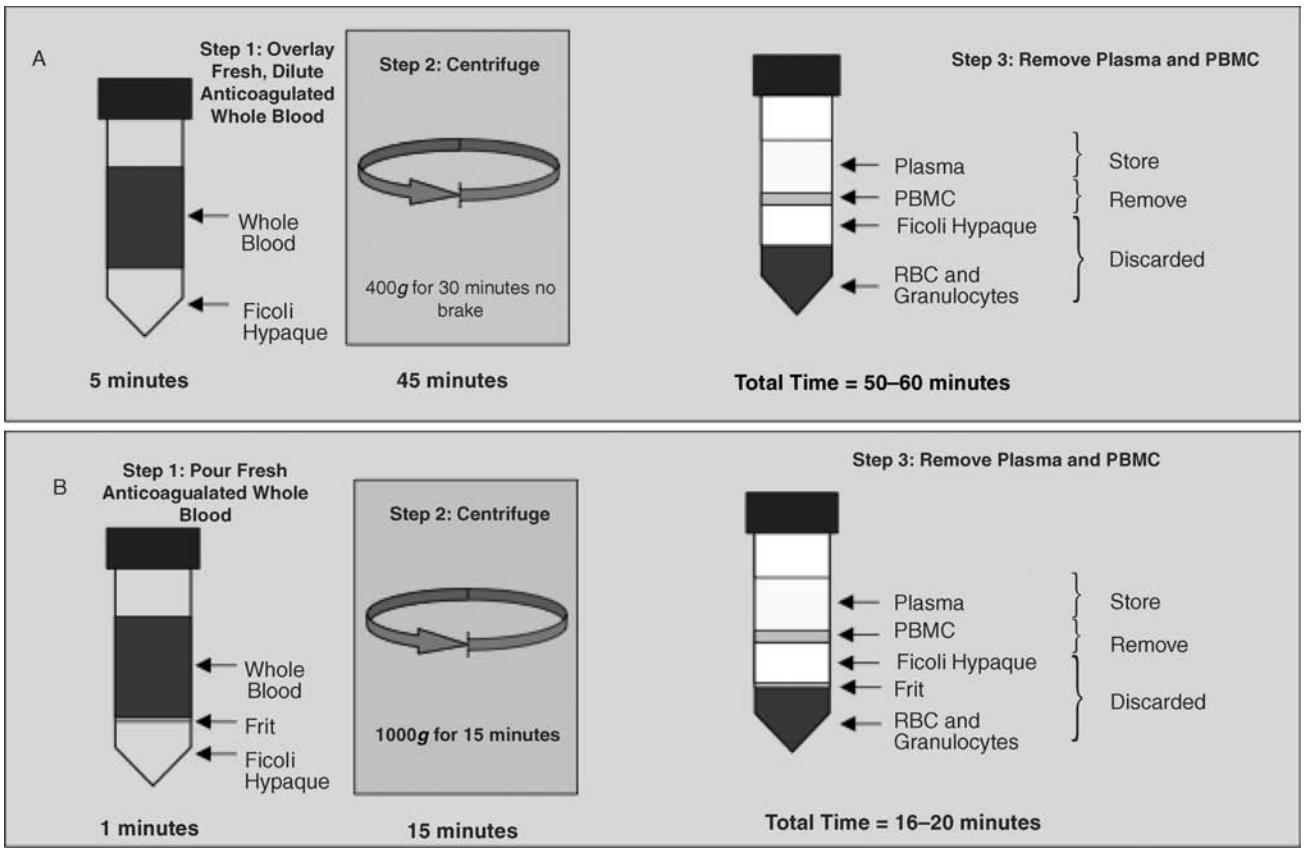


Figure 3. Gradient separation of peripheral blood mononuclear lymphocytes (PBMC). (a) The standard ficoll gradient method. (b) The Accuspin or Leucosep method. RBC = red blood cells, g = gravity. (Graphic courtesy of Greg Khoury and Clive Gray, National Institute for Communicable Diseases, Johannesburg, South Africa.)

may macroscopically observe the presence of gel spheres in the cellular layer, which are very difficult to distinguish from the actual PBMC. This has been observed after storage at temperatures $> 25\,^{\circ}$C. Where possible, the tubes should be stored at no $> 25\,^{\circ}$C. Once the tubes have blood drawn into them, an attempt should be made to keep them at temperatures of between 18 and 25 °C. Blood filled CPT should under no circumstances be stored on ice or next to an ice pack. It is recommended that they are separated from any ice-packs by bubble wrap or other type of insulation within a cooler so that the temperature fluctuations are kept to a minimum.

2. **Method:** (a) Specimens should be transported to the laboratory as soon as possible after collection. The manufacturer recommends the initial centrifugation to separate the lymphocytes be within 2 h. The samples may then be mixed by inversion and the processing completed preferably within 8 h after centrifugation. If there is a significant time delay, the specimens should be put into a cooler box and transported at room temperature (18–25 °C). (b) Spin tubes at room temperature (18–25 °C) in a horizontal rotor (swinging bucket head) for a minimum of 20 min and maximum of 30 min at $\sim 400\,g$. The brake is left off to assure that the PBMC layer is not jarred or disturbed while the centrifuge rotors are being mechanically halted. (c) Remove the tubes from the centrifuge and pipete the entire contents of the tube above the gel into a 50 mL tube. This tube will now contain both PBMC and undiluted plasma. An additional centrifugation step will allow removal of undiluted plasma if desired. Wash each CPT tube with 5 mL of PBS/1% Penicillin/Streptomycin (Pen/Strep). This wash step will remove cells from the top of the gel plug. Combine with cells removed from tube. This wash increases yield of cells by as much as 30–40%. (d) Spin down this tube at $300\,g$ for 15–20 min at room temperature with the brake on. (e) The PBMC pellet is resuspended in RPMI, 2% FBS and washed one more time to remove contaminating platelets. The PBMC are counted and cryopreserved or used as required.

3. **Separation of PBMC Using Accuspin or Leucosep Tubes.** More recently, the Leucosep and Accuspin tube have become available. Further information on the Leucosep is available at www.gbo.com and for the Accuspin at www.sigmaaldrich.com The principle of these tubes is the

same. The tube is separated into two chambers by a porous barrier made of highly transparent polypropylene (the frit). This biologically inert barrier allows elimination of the laborious overlaying of the sample material over Ficoll. The barrier allows separation of the sample material added to the top from the separation medium (ficoll added to the bottom). Figure 3 shows a comparison of the standard ficoll method and the Accuspin or Leucosep method. The tubes are available in two sizes and may be purchased with or without Ficoll. There is an advantage of buying the tubes without Ficoll because they can be stored at room temperature rather than refrigerated. This may be an important problem if cold space is limiting or cold chain is difficult. The expiration date of the Ficoll will not affect the tube expiration. The following procedure describes the separation procedure for Leucosep tubes that are not prefilled with Ficoll-hypaque. The Accuspin procedure is virtually identical. Note that whole blood can be diluted 1:2 with balanced salt solution. While this dilution step is not necessary, it can improve the separation of PBMC and enhance PBMC yield. The procedure is carried out using aseptic technique.

Protocol 2: Separation of PBMC Using Accuspin or Leucosep Tubes

1. Warm-up the separation medium (Ficoll-hypaque) to room temperature protected from light.
2. Fill the Leucosep tube with separation medium: 3 mL for the 14 mL tube and 15 mL for the 40 mL tube.
3. Close the tubes and centrifuge at 1000 *g* for 30 s at room temperature.
4. Pour the whole blood or diluted blood into the tube: 3–8 mL for the 14 mL tube and 15–30 mL for the 50 mL tube.
5. Centrifuge for 10 min at 1000 *g* or 15 min at 800 *g* in a swinging bucket rotor, with the centrifuge brake off. The brake is left off to assure that the PBMC layer is not jarred or disturbed while the centrifuge rotors are being mechanically halted.
6. After centrifugation, the sequence of layers from top to bottom should be plasma and platelets; enriched PBMC fraction; Separation medium; porous barrier; Separation medium; Pellet (erythrocytes and granulocytes).
7. Plasma can be collected to within 5–10 mm of the enriched PBMC fraction and further processed or stored for additional assays.
8. Harvest the enriched PBMC and wash with 10 mL of PBS containing 1% Pen/Strep and centrifuge at 250 *g* for 10 min.
9. The PBMC pellet is resuspended in RPMI, 2% FBS and washed one more time to remove contaminating platelets. The PBMC are counted and cryopreserved or used as required.

1. **Specimen Rejection Criteria**

Incomplete or inaccurate specimen identification.

Inadequate volume of blood in additive tubes (i.e., partially filed coagulation tube) can lead to inappropriate dilution of addition and blood.

Hemolysis (i.e., potassium determinations)

Specimen collected in the wrong tube (i.e., end product is serum and test requires plasma).

Improper handling (i.e., specimen was centrifuged and test requires whole blood).

Insufficient specimen or quantity not sufficient (QNS).

For PBMC, the rejection criteria are not usually evaluated at the time of draw due to the complexity of the tests performed. However, a minimum of 95% viability would be expected after PBMC separation unless the specimens have been subjected to heat or other adverse conditions (see note below).

The optimal time frame between collection of blood sample to processing, separation and cryopreservation of PBMC should be < 8 h or on the same day as collection. It is not always feasible to process, separate and cryopreserve PBMC within 8 h when samples are being shipped to distant processing centers. Under these conditions, PBMC left too long in the presence of anticoagulants or at noncompatible temperatures, adversely affect PBMC function and causes changes which affect the PBMC separation process (11).

There have been significant revisions to the procedures for the handling and processing of blood specimens; specimens for potassium analysis should not be recentrifuged because centrifugation may cause results to be falsely increased; the guidelines recommend that serum or plasma exposed to cells in a blood-collection tube prior to centrifugation should not exceed 2 h; storage recommendations for serum–plasma may be kept at room temperature up to 8 h, but for assays not completed within 8 h, refrigeration is recommended (2–8 °C), if the assay is not completed within 48 h serum–plasma should be frozen at or below −20 °C.

2. **Disclaimer.** The opinions or assertions contained herein are the private views of the author, and are not to be construed as official, or as reflecting true views of the Department of the Army or the Department of Defense.

BIBLIOGRAPHY

Cited References

1. McCall RE, Tankersley CM. Phlebotomy Essentials. Philadelphia: J.B. Lippincott; 1993.
2. Ernst DJ, Szamosi DI. 2005. Medical Laboratory Observer Clinical Laboratory, Specimen-collection standards complete major revisions, Available at www.mlo-online.com, Accessed 2005 Feb.
3. Arkin CF, et al. Procedures for the Collection of Diagnostic Blood Specimens by Venipuncture; CLSI (NCCLS) Approved

Standard – 5th ed. H3-A5, Vol 23, Number 32. NCCLS, Wayne (PA).
4. Becton, Dickinson and Company. 2004, BD Vacutainer Order of Draw for Multiple Tube Collections. Available at www.bd.com/vacutainer.
5. Centers for Disease Control and Prevention (CDC). Revised Guidelines for performing CD4+ T-cell determinations in persons infected with human immunodeficiency virus (HIV). MMWR. 1997;46(No.RR-2):1–29.
6. Deems D, et al. 1994, FACSCount White paper. Becton Dickenson. Available at www.bdbiosciences.com/immunocytometry systems/ whitepapers/pdf/FcountWP.pdf.
7. Dieye TN, et al. Absolute CD4 T-Cell Counting in Resource-Poor Settings: Direct Volumetric Measurements Versus Bead-Based Clinical Flow Cytometry Instruments. J Acquir Immune Defic Syndr 2005;39:32–37.
8. Maino VC, Maecker HT. Cytokine flow cytometry: a multiparametric approach for assessing cellular immune responses to viral antigens. Clin Immunol. 2004;110:222–231.
9. Wiseman JD et al. Procedures for the Handling and Processing of Blood Specimens; CLSI (NCCLS) Approved Guideline. 2nd ed. H18-A2. Vol. 19 Number 21. 1999. NCCLS, Wayne, PA
10. Boyum A. Isolation of mononuclear cells and granulocytes from human blood. Scand J Clin Lab Invest 1968;21:77–89.
11. Cox J et al. Accomplishing cellular immune assays for evaluation of vaccine efficacy. In: Hamilton RG, Detrich B, Rose NR; Manual Clinical Laboratory Immunology 6th ed. Washington (DC): ASM Press; 2002. Chapt. 33. pp 301–315.

See also ANALYTICAL METHODS, AUTOMATED; CELL COUNTERS, BLOOD; DIFFERENTIAL COUNTS, AUTOMATED.

BLOOD FLOW. See BLOOD RHEOLOGY; HEMODYNAMICS.

BLOOD GAS MEASUREMENTS

AHMAD ELSHARYDAH
RANDALL C. CORK
Louisiana State University
Shreveport, Louisiana

INTRODUCTION

Blood gas measurement–monitoring is essential to monitor gas exchange in critically ill patients in the intensive care units (1,2), and "standard of care" monitoring to deliver general anesthesia (3). It is a cornerstone in the diagnosis and management of the patient's oxygenation and acid–base disorders (4). Moreover, it may indicate the onset or culmination of cardiopulmonary problems, and may help in evaluating the effectiveness of the applied therapy. Numerous studies and reports have shown the significance of utilizing blood gas analyses in preventing serious oxygenation and acid–base problems. This article gives a summarized explanation of the common methods and instruments used nowadays in blood gas measurements in clinical medicine. This explanation includes a brief history of the development of these methods and instruments, the principles of their operation, a general description of their designs, and some of their clinical uses, hazards, risks, limitations, and finally the direction in the future to improve these instruments or to invent new ones. Blood gas measurement in clinical medicine can be classified into two major groups: (1) Noninvasive blood gas measurement, which includes blood oxygen–carbon dioxide measurement–monitoring by using different types of pulse oximeters (including portable pulse oximeters), transcutaneous oxygen partial pressure–carbon dioxide partial pressure (PO_2/PCO_2) monitors, intrapartum fetal pulse oximetry, cerebral oximetry, capnometry, capnography, sublingual capnometry, and so on; (2) invasive blood gas measurement, which involves obtaining a blood sample to measure blood gases by utilizing blood gas analyzers (in a laboratory or by using a bedside instrument), or access to the vascular system to measure/monitor blood gases. Examples include, but not limited to, mixed venous oximetry (SvO_2) monitoring by utilizing pulmonary artery catheter or jugular vein (SvO_2) measurement (5); continuous fibroptic arterial blood gas monitoring, and so on. In this article, we will talk about some of these methods; others have been mentioned in other parts of this encyclopedia.

BASIC CONCEPTS IN INVASIVE AND NONINVASIVE BLOOD GAS MEASUREMENTS

The Gas Partial Pressure

Gases consist of multiple molecules in rapid, continuous, random motion. The kinetic energy of these molecules generate a force as the molecules collide with each other and bounce from one surface to another. The force per unit area of a gas is called pressure, and can be measured by a device called a manometer. In a mixture of gases (e.g., a mixture of O_2, CO_2, and water vapor), several types of gas molecules are present within this mixture, and each individual gas (e.g., O_2 or CO_2) in the mixture is responsible for a portion of the total pressure. This portion of pressure is called partial pressure (P). According to Dalton's law, the total pressure is equal to the sum of partial pressures in a mixture of gases. Gases dissolve freely in liquids, and may or may not react with the liquid, depending on the nature of the gas and the liquid. However, all gases remain in a free gaseous phase to some extent within the liquid. Gas dissolution in liquids is a physical, not chemical, process. Therefore, gases (e.g., CO_2, O_2) dissolved in liquid (blood) exist in two phases: liquid and gaseous phase. Henry's law states that the partial pressure of a gas in the liquid phase equilibrates with the partial pressure of that gas in the gaseous phase (6,7).

BLOOD GAS ELECTRODES

Basic Electricity Terms

Electricity is a form of energy resulting from the flow of electrons through a substance (conductor). Those electrons flow from a negatively charged pole called Cathode, which has an excess of stored electrons, to a positively charged pole called Anode, which has a relative shortage

of electrons. The potential is the force responsible for pumping these electrons between the two poles. The greater the difference in electron concentration between these two poles, the greater is the potential. Volt is the potential measurement unit. The electrical current is the actual flow of electrons through a conductor. Ampere (amp) is the unit of measurement for the electrical current. Conductors display different degree of electrical resistance to the flow of the electrical current. The unit of the electrical resistance is ohm (Ω). Ohm's law states: voltage = current × resistance.

The Principles of Blood Gas Electrodes

Blood gas electrodes are electrochemical devices used to measure directly pH and blood gases. These blood gas electrodes use electrochemical cells. The electrochemical cell is an apparatus that consists of two electrodes placed in an electrolyte solution. These cells usually incorporated together (one or more cells) to form an electrochemical cell system. These systems are used to measure specific chemical materials (e.g., PO_2, PCO_2 and pH). The basic generic blood gas electrode consists of two electrode terminals, which are also called half-cells: one is called the working half-cell where the actual chemical analysis occurs, or electrochemical change is taken place; and the other one is called the reference half-cell. The electrochemical change occurring on the working terminal is compared to the reference terminal, and the difference is proportional to the amount of blood gas in the blood sample (6,7).

PO_2 Electrode

The PO_2 electrode basically consists of two terminals (1). The cathode, which usually made of platinum (negatively charged) and (2) the anode, which usually made of silver–sliver chloride (positively charged). How does this unit measure PO_2 in the blood sample? As shown in Fig. 1, the electricity source (battery or wall electricity) supplies the platinum cathode with energy (voltage of ~700 mV). This voltage attracts oxygen molecules to the cathode surface, where they react with water. This reaction consumes four electrons for every oxygen molecule reacts with water and produces four hydroxyl ions. The consumed four electrons, in turn, are replaced rapidly in the electrolyte solution as silver and chloride react at the anode. This continuous reaction leads to continuous flow of electrons from the anode to the cathode (electrical current). This electrical current is measured by using an ammeter (electrical current flow meter). The current generated is in direct proportion to the amount of dissolved oxygen in the blood sample, which in direct proportion to PO_2 in that sample.

Oxygen Polarography

The electrical current and PO_2 have a direct (linear) relationship when a specific voltage is applied to the cathode. Therefore, a specific voltage must be identified, to be used in PO_2 analysis. The polarogram is a graph that shows the relationship between voltage and current at a constant PO_2. As shown in Fig. 2, when the negative voltage applied to the cathode is increased, the current increases initially, but soon it becomes saturated. In this plateau region of the polarogram, the reaction of oxygen at the cathode is so fast that the rate of reaction is limited by the diffusion of oxygen to the cathode surface. When the negative voltage is further increased, the current output of the electrode increases rapidly due to other reactions, mainly, the reduction of water to hydrogen. If a fixed voltage in the plateau region (e.g., −0.7 V) is applied to the cathode, the current output of the electrode can be linearly calibrated to the dissolved oxygen. Note that the current is proportional not to the actual concentration, but to the activity or equivalent partial pressure of dissolved oxygen. A fixed voltage between −0.6 and −0.8 V is usually selected as the

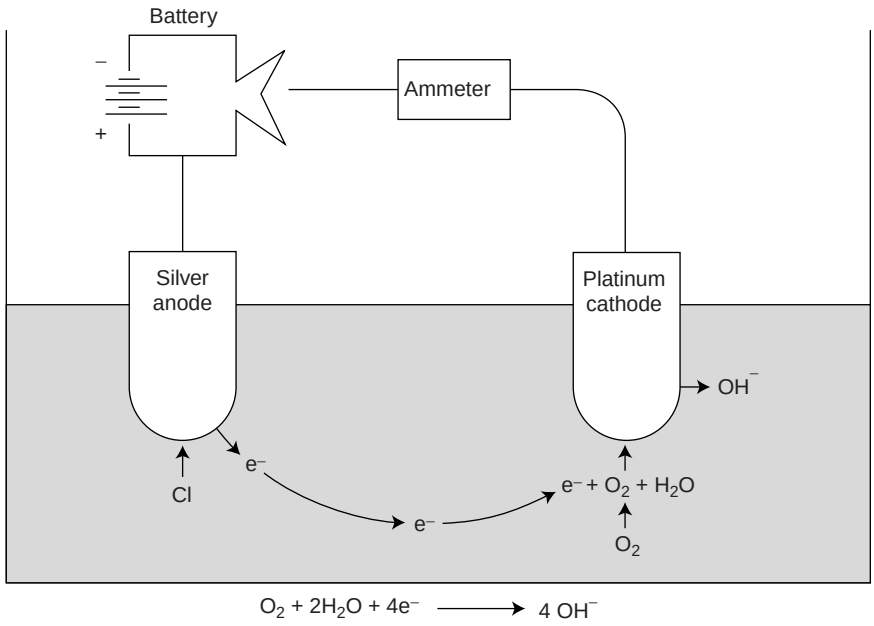


Figure 1. PO_2 electrode.

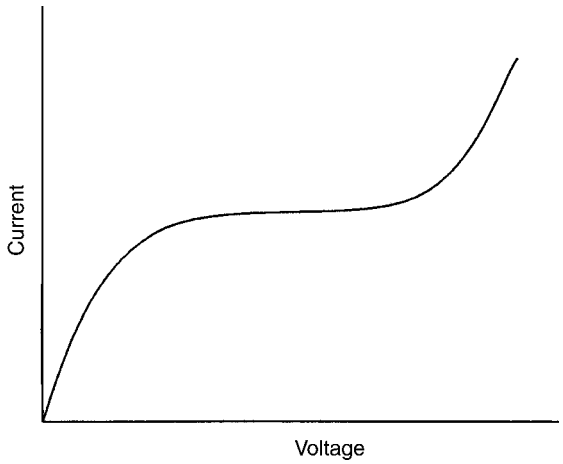


Figure 2. Polarogram.

polarization voltage when using Ag/AgCl as the reference electrode.

pH Electrode

The pH electrode uses voltage to measure pH, rather than actual current as in PO_2 electrode. It compares a voltage created through the blood sample (with unknown pH) to known reference voltage (in a solution with known pH). To make this possible, the pH electrode basically needs four electrode terminals (Fig. 3), rather than two terminals (as in the PO_2 electrode). Practically, one common pH-sensitive glass electrode terminal between the two solutions is adequate. This glass terminal allows the hydrogen ions to diffuse into it from each side. The difference in the hydrogen ions concentration across this glass terminal creates a net electrical potential (voltage). A specific equation is used to calculate the blood sample pH, using the reference fluid pH, the created voltage, and the fluid temperature.

PCO_2 Electrode

The PCO_2 electrode is a modified pH electrode. There are two major differences between this electrode and the pH electrode. The first difference is that in this electrode, the blood sample comes in contact with a CO_2 permeable membrane (such as Teflon, Silicone rubber), rather than a pH-sensitive glass (in the pH electrode), as shown in Fig. 4. The CO_2 from the blood sample diffuses via the CO_2 permeable (silicone) membrane into a bicarbonate solution. The amount of the hydrogen ions produced by the hydrolysis process in the bicarbonate solution is proportional to the amount of the CO_2 diffused through the silicone membrane. The difference in the hydrogen ions concentration across the pH-sensitive glass terminal creates a voltage. The measured voltage (by voltmeter) can be converted to PCO_2 units. The other difference is that the CO_2 electrode has two similar electrode terminals (silver–silver chloride). However, the pH electrode has two different electrode terminals (silver–silver chloride and mercury–mercurous chloride).

BLOOD GAS PHYSIOLOGY (8,9)

Oxygen Transport

Oxygen is carried in the blood in two forms: A dissolved small amount and a much bigger, more important component combined with hemoglobin. Dissolved oxygen plays a small role in oxygen transport because its solubility is so low, 0.003 mL O_2/100 mL blood per mmHg (133.32 Pa). Thus, normal arterial blood with a PO_2 of $\sim$100 mmHg (13332.2 Pa) contains only 0.3 mL of dissolved oxygen per 100 mL of blood, whereas $\sim$20 mL is combined with hemoglobin. Hemoglobin consists of heme, an iron–porphyrin compound, and globin, a protein that has four polypeptide chains. There are two types of chains, alpha and beta, and differences in their amino acid sequences give rise to different types of normal and abnormal human hemoglobin, such as, hemoglobin F (fetal) in the newborn, and hemoglobin S in the sickle cell anemia patient. The combination of oxygen (O_2) with hemoglobin (Hb) (to form oxyhemoglobin–HbO_2) is an easily reversible. Therefore, blood is able to transport large amounts of oxygen.

The relationship between the partial pressure of oxygen and the number of binding sites of the hemoglobin that have oxygen attached to it, is known as the oxygen dis-

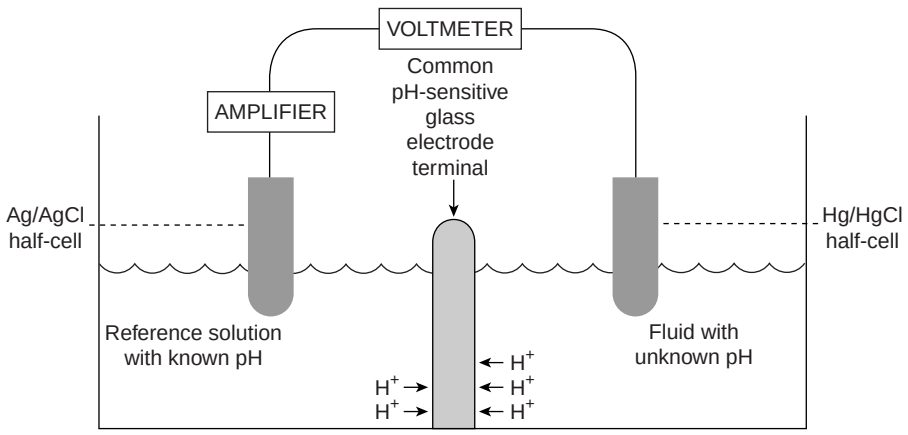


Figure 3. pH electrode.

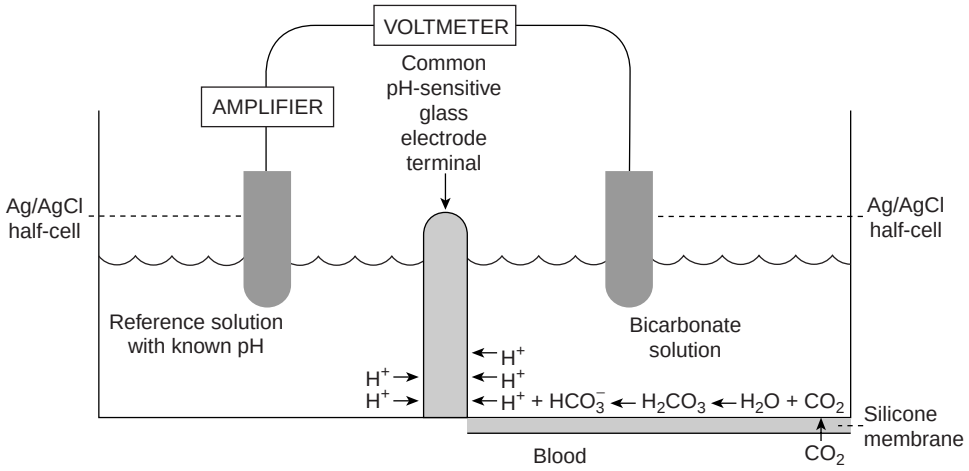


Figure 4. PCO_2 electrode.

sociation curve (Fig. 5). Each gram of pure hemoglobin can combine with 1.39 mL of oxygen, and because normal blood has ~15 g Hb/100 mL, the oxygen capacity (when all the binding sites are full) is ~20.8 mL O_2/100 mL blood. The total oxygen concentration of a sample of blood, which includes the oxygen combined with Hb and the dissolved oxygen, is given by $(Hb \times 1.36 \times SaO_2) + (0.003 \times PaO_2)$ Hb is the hemoglobin concentration.

The characteristic shape of the oxygen dissociation curve has several advantages. The fact that the upper portion is almost flat means that a fall of 20–30 mmHg in arterial PO_2 in a healthy subject with an initially normal value (e.g., ~100 mmHg or 13332.2 Pa) causes only a minor reduction in arterial oxygen saturation. Another consequence of the flat upper part of the curve is that loading of oxygen in the pulmonary capillary is hastened. This results from the large partial pressure difference between alveolar gas and capillary blood that continues to exist even when most of the oxygen has been loaded. The steep lower part of the oxygen dissociation curve means that considerable amounts of oxygen can be unloaded to the peripheral tissues with only a relatively small drop in capillary PO_2. This maintains a large partial pressure difference between the blood and the tissues, which assists in the diffusion process. Various factors affect the position of the oxygen dissociation curve, as shown in Fig. 5. It is shifted to the right by an increase of temperature, hydrogen ion concentration, PCO_2, and concentration of 2,3-diphosphoglycerate in the red cell. A rightward shift indicates that the affinity of oxygen for hemoglobin is reduced. Most of the effect of the increased PCO_2 in reducing the oxygen affinity is due to the increased hydrogen concentration. This is called the Bohr effect, and it means that as peripheral blood loads carbon dioxide, the unloading of oxygen is assisted. A useful measure of the position of the dissociation curve is the PO_2 for 50% oxygen saturation; this is known as the $P50$. The normal value for human blood is ~27 mmHg (3599.6 Pa).

Carbon Dioxide Transport

Carbon dioxide is transported in the blood in three forms: dissolved, as bicarbonate, and in combination with proteins such as carbamino compounds (Fig. 6). Dissolved carbon dioxide obeys Henry's law (as mentioned above). Because carbon dioxide is some 24 times more soluble than oxygen in blood, dissolved carbon dioxide plays a much more significant role in its carriage compared to oxygen. For example, ~10% of the carbon dioxide that evolves into the alveolar gas from the mixed venous blood comes from the dissolved form. Bicarbonate is formed in blood by the following hydration reaction:

$$CO_2 + H_2O \leftrightarrow H_2CO_3 \leftrightarrow H^+ + HCO_3^-$$

The hydration of carbon dioxide to carbonic acid (and vice versa) is catalyzed by the enzyme carbonic anhydrase (CA), which is present in high concentrations in the red cells, but is absent from the plasma. However, some carbonic anhydrase is apparently located on the surface of the endothelial cells of the pulmonary capillaries. Because of the presence of carbonic anhydrase in the red cell, most of the hydration of

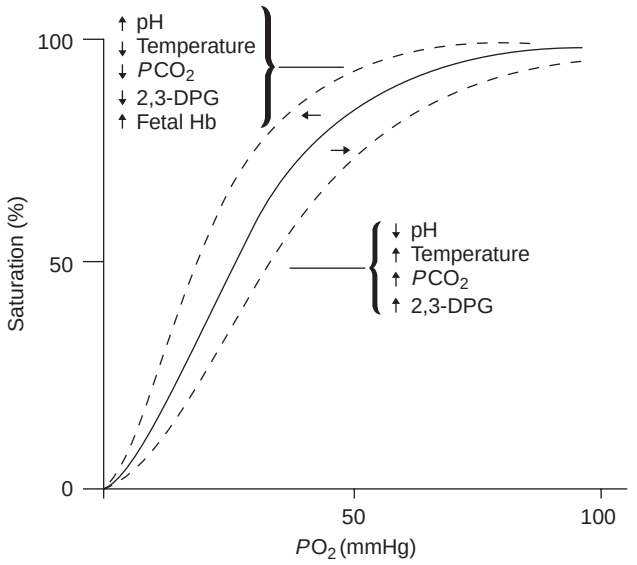


Figure 5. Oxygen dissociation curve and the effects of different factors on it.

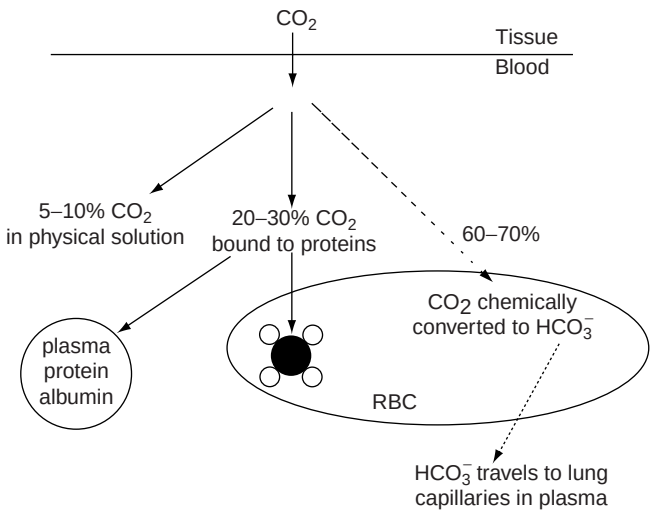


Figure 6. Carbon dioxide transport in blood.

carbon dioxide occurs there, and bicarbonate ion moves out of the red cell to be replaced by chloride ions to maintain electrical neutrality (chloride shift). Some of the hydrogen ions formed in the red cell are bound to Hb, and because reduced Hb is a better proton acceptor than the oxygenated form, deoxygenated blood can carry more carbon dioxide for a given PCO_2 than oxygenated blood can. This is known as the Haldane effect. Carbamino compounds are formed when carbon dioxide combines with the terminal amine groups of blood proteins. The most important protein is the globin of hemoglobin. Again, reduced hemoglobin can bind more carbon dioxide than oxygenated hemoglobin, so the unloading of oxygen in peripheral capillaries facilitates the loading of carbon dioxide, whereas oxygenation has the opposite effect. The carbon dioxide dissociation curve, as shown in Fig. 7, is the relationship between PCO_2 and total carbon dioxide concentration. Note that the curve is much more linear in its working range than the oxygen dissociation curve, and also that, as we have seen, the lower the saturation of hemoglobin with oxygen, the larger the carbon dioxide concentration for a given PCO_2.

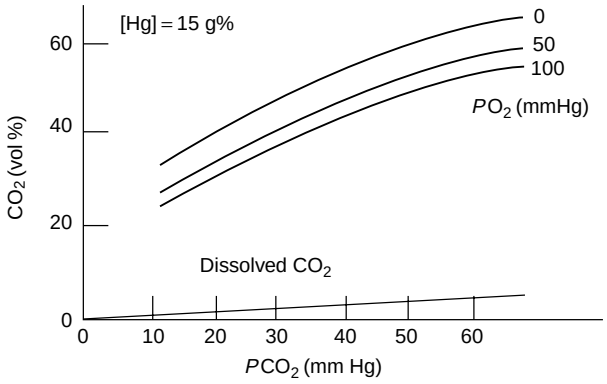


Figure 7. The carbon dioxide dissociation curve showing the effect of PO_2 variations.

OXIMETRY

Historical Development

Oximetry has its origins in the early 1860s (10), when Felix Hoppe-Seyler described the hemoglobin absorption of light using the spectroscope. He demonstrated that the light absorption was changed when blood was mixed with oxygen, and that hemoglobin and oxygen formed a compound called oxyhemoglobin. Soon after, George Gabriel Stokes reported that hemoglobin was in fact the carrier of oxygen in the blood. In 1929, Glen Allan Millikan (11), an American physiologist, began construction of a photoelectric blood oxygen saturation meter, which, used to measure color changes over time when desaturated hemoglobin solutions were mixed with oxygen solutions in an experimental setting. The use of photoelectric cells later proved to be crucial to the development of oximeters. In 1935, Kurt Kramer demonstrated, for the first time, *in vivo* measurement of blood oxygen saturation in animals. The same year, Karl Matthes introduced the ear oxygen saturation meter. This was the first instrument able to continuously monitor blood oxygen saturation in humans. In 1940, J.R. Squire introduced a two-channel oximeter that transmitted red and infrared (IR) light through the web of the hand. In 1940, Millikan and colleagues developed a functioning oximeter, and introduced the term "oximeter" to describe it. The instrument used an incandescent, battery-operated light and red and green filter. In 1948, Earl Wood of the Mayo Clinic made several improvements to Millikan's oximeter, including the addition of a pressure capsule. Then, in the 1950s, Brinkman and Zijlstra of the Netherlands developed the reflectance oximetry. However, oximetry did not fully achieve clinical applicability until the 1970s.

Principles of Operation

It is important to understand some of the basic physics principles that led to the development of oximetry and pulse oximetry. This is a summary of these different physics principles and methods (7,12).

Spectrophotometry. The spectroscope is a device which was used initially to measure the exact wavelengths of light emitted from a light generator (bunsen burner) (10). Each substance studied with the spectroscope has its unique light emission spectrum, in other words, each substance absorbed and then emitted light of different wavelengths. The graph of the particular pattern of light absorption–emission of sequential light wavelengths called the absorption spectrum. Figure 8 reveals the absorption spectra of common forms of hemoglobin.

Colorimetry. Colorimetry is another method of qualitative analysis (10). In this method, the color of known substance is compared of that of unknown one. This method is not highly exact, because it depends on visual acuity and perception.

Photoelectric Effect. The photoelectric effect is the principle behind spectrophotometry. It is defined as the ability of light to release electrons from metals in proportion to the

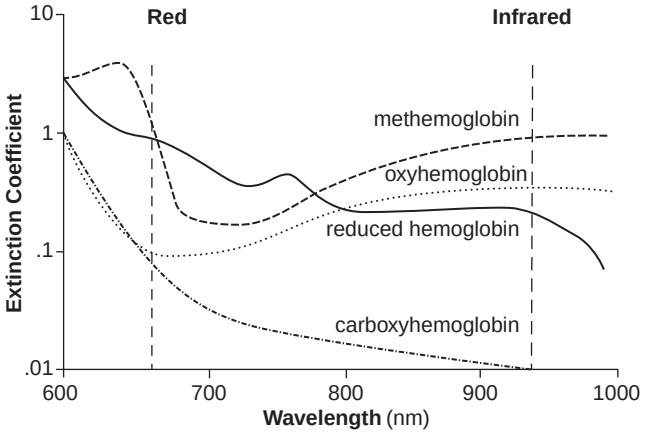


Figure 8. Absorption spectra of common forms of hemoglobin. Absorption spectra of oxyhemoglobin, deoxyhemoglobin, methemoglobin, carboxyhemoglobin.

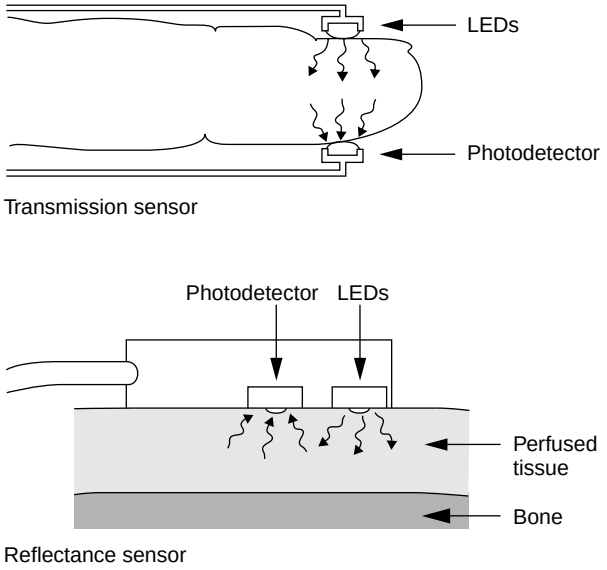


Figure 9. Major components of transmission and reflection oximeters.

intensity of the light. In spectrophotometry, light passes via a filter that converts the light into a specific wavelength. This light then passes through a container that contains the substance being analyzed. This substance absorbs part of this light and emits the remaining part, which goes through a special cell. This cell is connected to a photodetector, which detects and measures the emitting light (spectrophotometry). This method can be used for quantitive as well as qualitative analyses.

Lambert–Beer Law. This law combines the different factors that affect the light absorption of a substance:

$$\log 10 I_o/I_x = kcd$$

I_o = intensity of light incident on the specimen

I_x = intensity of the transmitted light

I_o/I_x = optical density

As shown in the above formula, the concentration of absorbing substance, the path length of the absorbing medium (d) and the characteristics of the substance and the light wavelength (k = constant) all affect light absorption (12).

Transmission Versus Reflection Oximetry. When the light at a particular wavelength passes through a blood sample, which contains Hb, this light would be absorbed, transmitted, or reflected. The amount of the absorbed, transmitted, or reflected light at those particular wavelengths is determined by various factors, including the concentration (Lambert–Beer law) and the type of the Hb present in the blood sample. The amount of light transmitted through the blood sample at a given wavelength is related inversely to the amount of light absorbed or reflected. The transmission oximetry is a method to determine the arterial oxygen saturation (S_aO_2) value by measuring the amount of light transmitted at certain wavelengths. On the other side, in the reflection oximetry, measuring the amount of light reflected is used to determine the S_aO_2 value. The significant difference between these two methods is the location of the photodetecor (Fig. 9). In the reflection method, the photodetector is on the same side of the light source. However, in the transmission oximetry, it is on the opposite site side of the light source (7,12).

Oximetry Versus Cooximetry. Each form of hemoglobin (e.g., oxyhemoglobin, deoxygenated hemoglobin, carboxyhemoglobin, methemoglobin) has its own unique absorption–transmission–reflection spectrum. By plotting the relative absorbance to different light wavelengths for both oxyhemoglobin and deoxygenated Hb as shown in Fig. 10. It is clear that these two hemoglobins absorb light differently at different light wavelengths. This difference is big in some light wavelengths (e.g., 650 nm in the red region), and small or not existing in other light wavelengths. The isosbestic point (13) is the light wavelength at which there is no difference between these two hemoglobins in absorbing light ($\sim$805 nm near the IR region). The difference in these two wavelengths can be used to calculate the S_aO_2.

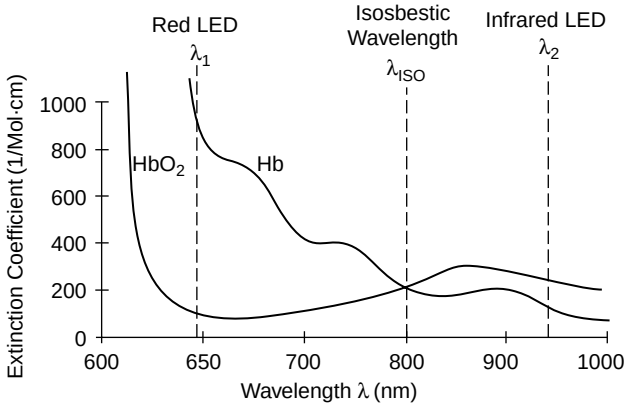


Figure 10. Light absorption spectra of oxygenated and deoxygenated hemoglobin.

However, these two hemoglobins are not only the hemoglobins exist in the patient's blood. There are other abnormal hemoglobins (dyshemoglobins) that can join these two hemoglobins in some abnormal conditions (such as carboxyhemoglobin and methemoglobin). Each one of these dyshemoglobins has its unique transmission–reflection–absorption spectrum. Some of these spectra are very close to the oxyhemoglobin spectrum at the routinely used two light wavelengths (see above). This makes these two wavelengths are incapable in detecting those dyshemoglobins. Therefore, the use of regular oximeters in these conditions may lead to erroneous and false readings, which may lead to detrimental effects on the patient's care. To overcome this significant problem a special oximeter (cooximeter, i.e., cuvette oximeter) is needed when there is a suspicion of presence of high level of dyshemoglobins in the patient's blood. Functional S_aO_2 is the percentage of oxyhemoglobin compared to sum of oxy- and deoxyhemoglobins. Therefore, the abnormal hemoglobins are not directly considered in the measurement of functional S_aO_2 by using regular oximetry. Cooximetry uses four or more light wavelengths, and has the ability to measure carboxyhemoglobin and methemoglobin as well as normal hemoglobins. The fractional S_aO_2 measures the percentage of oxyhemoglobin to all hemoglobins (normal and abnormal) present in the blood sample (14,15).

EAR OXIMETRY

Historical Development

In 1935, Matthes (16,17) showed that transmission oximetry could be applied to the external ear. However, a major problem with noninvasive oximetry applied to the ear was the inability to differentiate light absorption due to arterial blood from that due to other ear tissue and blood. In the following years, two methods were tried to solve this problem. The first was increasing local perfusion by heating the ear, applying vasodilator, or rubbing the ear. The second was comparing the optical properties of a "bloodless" earlobe (by compressing it using a special device) to the optical properties of the perfused ear lobe. Arterial S_aO_2 was then determined from the difference in these different measurements. This step was a significant step toward an accurate noninvasive measurement of S_aO_2. In 1976, Hewlett-Packard (18) used the collected knowledge about ear oximetry to that date to develop the model 47201A ear oximeter, Fig. 11.

HEWLETT-PACKARD EAR OXIMETER

This oximeter (18) is based on the measured light transmission at eight different wavelengths, which made this sensor less accurate and more complex than pulse oximeters. It used a high intensity tungsten lamp that generated a broad spectrum of light wave lengths. This light passes through light filters, then enters a fiberoptic cable, which carrys the filtered light to the ear. A second fibroptic cable carries the light pulses transmitted through the ear to the device for detection and analysis. The ear probe is relatively bulky ($\sim 10 \times 10$ cm) equipped with a temperature-controlled heater (to keep temperature of 41°C). It is attached to the antihelix after the ear has been rubbed briskly. This monitor is no longer manufactured because of its bulkiness and cost, and because of the development widely of a more accurate, smaller, and cost-effective monitor, the pulse oximeter.

Figure 11. The Hewlett-Packard Model 47201A ear oximeter.

PULSE OXIMETRY

Historical Development

In the early 1970s, Takuo Aoyagi (16,19,20), a Japanese physiological bioengineer, introduced pulse oximetry, the underlying concept of which had occurred to him while trying to cancel out the pulsatile signal of an earpiece densitometer with IR light. In early 1973, Dr. Susumu Nakajima, a Japanese surgeon, learned of the idea and ordered oximeter instruments from Nihon Kohden. After several prototypes were tested, Aoyagi and others delivered the first commercial pulse oximeter in 1974. This instrument was the OLV-5100 ear pulse oximeter, (Fig. 12). In 1977, the Minolta Camera Company

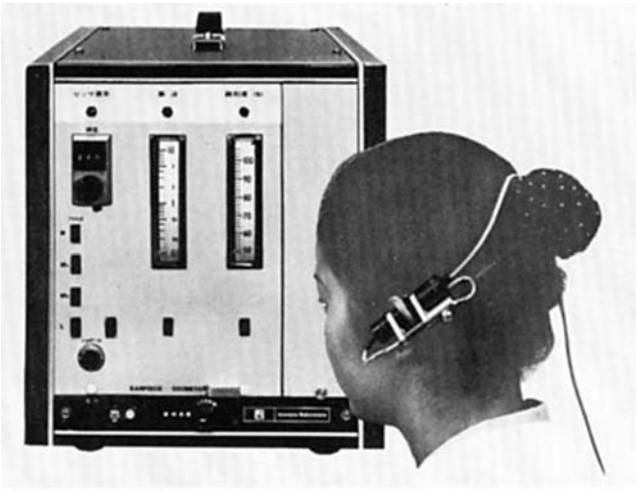


Figure 12. The OLV-5100 ear pulse oximeter, the first commercial pulse oximeter, it was introduced by Nihon Kohden in 1974.

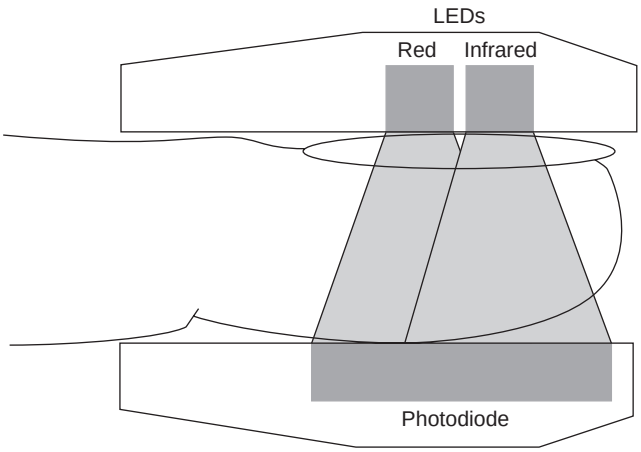


Figure 13. The basic components of a pulse oximeter sensor. Two LEDs with different wavelengths as light sources and a photodiode as receiver.

introduced the Oximet MET-1471 pulse oximeter with a fingertip probe and fiberoptic cables. Nakajima and others tested the Oximet MET-1471 and reported on it in 1979. In the years since, pulse oximetry has become widely used in a number of fields, including Anesthesia, intensive care, and neonatal care.

Principles of Operation

Pulse oximetry differs from the previously described oximetry in that it does not rely on absolute measurements, but rather on the pulsations of arterial blood. Oxygen saturation is determined by monitoring pulsations at two wavelengths and then comparing the absorption spectra of oxyhemoglobin and deoxygenated hemoglobin (20,21). Pulse oximetry uses a light emitter with red and infrared LEDs (light-emitting diodes) that shine through a reasonably translucent site with good blood flow (Fig. 13). Typical adult–pediatric sites are the finger, toe, pinna (top), or lobe of the ear. Infant sites are the foot or palm of the hand and the big toe or thumb. On the opposite side of the emitter is a photodetector that receives the light that passes through the measuring site. There are two methods of sending light through the measuring site (see above) (Fig. 9). The transmission method is the most common type used, and for this discussion the transmission method will be implied. After the transmitted red (R) and IR signals pass through the measuring site and are received at the photodetector, the R/IR ratio is calculated. The R/IR is compared to a "look-up" table (made up of empirical formulas) that converts the ratio to pulse oxygen saturation (S_pO_2) value. Most manufacturers have their own tables based on calibration curves derived from healthy subjects at various S_pO_2 levels. Typically, an R/IR ratio of 0.5 equates to approximately 100% S_pO_2, a ratio of 1.0 to $\sim$82% S_pO_2, while a ratio of 2.0 equates to 0% S_pO_2. The major change that occurred from the eight-wavelength Hewlett-Packard oximeters (see above) of the 1970s to the oximeters of today was the inclusion of arterial pulsation to differentiate the light absorption in the measuring site due

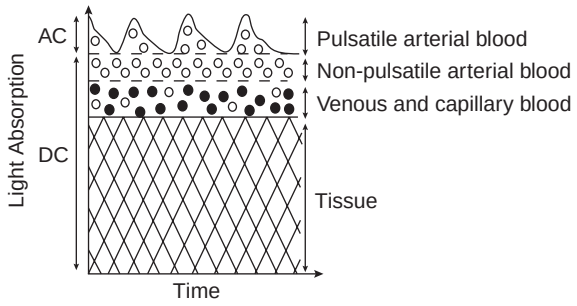


Figure 14. Schematic Representation of light absorption in adequately perfused tissue.

to skin, tissue, and venous blood from that of arterial blood. At the measuring site there are several light absorbers (some of them are constant) such as skin, tissue, venous blood, and the arterial blood (Fig. 14). However, with each heart beat the heart contracts and there is a surge of arterial blood, which momentarily increases arterial blood volume across the measuring site. This results in more light absorption during the surge. Light signals received at the photodetector are looked at as a waveform (peaks with each heartbeat and troughs between heartbeats). If the light absorption at the trough, which should include all the constant absorbers, is subtracted from the light absorption at the peak, then the resultants are the absorption characteristics due to added volume of blood only, which is arterial blood. Since peaks occur with each heartbeat or pulse, the term "pulse oximetry" was applied.

New Technologies

Conventional pulse oximetry accuracy degrades during motion and low perfusion. This makes it difficult to depend on these measurements when making medical decisions. Arterial blood gas tests have been and continue to be commonly used to supplement or validate pulse oximeter readings. Pulse oximetry has gone through many advances and developments since the Hewlett-Packard Model 47201A ear oximeter invention in 1976. There are several types of pulse oximeters manufactured by different companies available in the market nowadays. Different technologies have been used to improve pulse oximetry quality and decrease its limitations, which would lead eventually to better patient care. Figure 15 shows a modern pulse oximeter (Masimo Rad-9) designed by Masimo using the Signal Extraction Technology (Masimo SET) (22,23), is a software system composed of five parallel algorithms designed to eliminate nonarterial "noise" in a patient's blood flow. This monitor display includes: S_pO_2, pulse rate, alarm, trend, perfusion index (PI) (24), signal IQ, and plethysmographic waveform. Moreover, Masimo manufactures a handheld pulse oximeter by utilizing the same technology (Masimo SET) as shown in Fig. 16. Its small size ($\sim 15.7 \times 7.6 \times 3.5$ cm) and broad catalog of features make it suited for hospital, transport, and home use. Nellcor (25) uses the OxiMax technology to produce a list of pulse oximetry monitors and sensors. These sensors have a small digital memory chip that transmits

Figure 15. Masimo Rad-9 pulse oximeter.

Figure 17. Nellcor N-595 pulse oximeter.

sensor-specific data to the monitor. These chips contain all the calibration and operating characteristics for that sensor design. This gives the monitor the flexibility to operate accurately with a diverse range of sensor designs without the need for calibrating each sensor to the specific monitors. This opens a new area of pulse oximetry innovations. Figure 17 reveals some of the Nellcor monitors and sensors available. Furthermore, Nellcor has designed a handheld pulse oximeter, as shown in Fig. 18, compatible with its line of OxiMax pulse oximetry sensors. Nellcor combines two advanced technologies in measuring blood gases: the OxiMax technology and Microstream CO_2 technology (see the section Capnography) to produce a S_pO_2 and end-tidal CO_2 partial pressure (PetCO_2) handheld capnograph–pulse oximeter to monitor both S_pO_2 and PetCO_2. Several additional parameters are now available on the modern oximeter, and they add additional functionality for these monitors and decrease their limitations. One of these parameters is called the "perfusion index" (PI) (24,26). This is a simple measure of the change that has occurred in the tissue-under-test (e.g., the finger) over the cardiac cycle. When this parameter was first recognized as being something that a pulse oximeter could measure, it was difficult to imagine a value to the measurement because it is affected by so many different physiological and environmental variables, including systemic vascular resistance, volume status, blood pressure, and ambient temperature. But as time continues to pass since its introduction, more applications for PI have been found. The most obvious use for perfusion index is as an aid in sensor placement. It provides a means to quantify the validity of a given sensor site and, where desired, to maximize measurement accuracy. Perfusion index has also provided a simple and easy way test for sufficient collateral blood flow in the ulnar artery to allow for harvest of the radial artery for coronary artery bypass graft (CABG) surgery and for monitoring peripheral perfusion in critically ill patients.

Figure 16. Masimo Rad-5 handheld pulse oximeter.

Clinical Uses

Pulse oximeters are widely used in clinical practice (27–30). They are used extensively in the intensive care units to monitor oxygen saturation, and to detect and prevent hypoxemia. Monitoring oxygen saturation during anesthesia is a standard of care, which is almost always done by pulse oximeters. Pulse oximeters are very helpful in monitoring patients during procedures like bronchoscopy, endoscopy, cardiac catheterization, exercise testing,

Figure 18. Nellcor N-45 handheld pulse oximeter.

Figure 19. Portable Nonin Onyx 9500 pulse oximeter.

and sleep studies. Also, they are commonly used during labor and delivery for both the mother and infant. These sensors have no significant complications related to their use. There are several types of portable pulse oximeters on the market. These oximeters are small in size, useful for patients transport, and can be used at home. Figure 19 shows one of these pulse oximeters.

Accuracy and Limitations

The accuracy of pulse oximeters in measuring exact saturation has been shown to be $\sim \pm 4\%$ as compared to blood oximetry measurements. Several studies have shown that with low numbers of S_aO_2, there is a decreased correlation between S_pO_2 and S_aO_2, especially when $S_aO_2 < 70\%$ and in unsteady conditions (31). However, newer technologies have improved accuracy during these conditions substantially. Another factor that influences the accuracy of pulse-oximetry is the response time. There is a delay between a change in S_pO_2 and the display of this change. This delay ranges from 10 to 35 s. Pulse-oximeters have several limitations that may lead to inaccurate readings. One of its most significant limitations is that it estimates the S_aO_2, not the arterial oxygen tension (P_aO_2). Another limitation is the difficulty these sensors have in detecting arterial pulsation in low perfusion states (low cardiac output, hypothermia etc.) (32). Furthermore, the presence of dyshemoglobins (e.g., methemoglobin, carboxyhemoglobin) (15) and diagnostic dyes (e.g., methylene blue, indocyanine green, and indigo carmine) (33) affects the accuracy of these monitors, leading to false readings. High carboxyhemoglobin levels will falsely elevate S_pO_2 readings, which may lead to a false sense of security regarding the patient's oxygenation, and possible disastrous outcome. CO-oximetry should be used to measure S_aO_2 in every patient who is suspect for elevated carboxyhemoglobin (such as fire victims). Methemoglobinemia may lead to false $\sim 85\%$ saturation reading. The clinician should be alert to the potential causes and possibility of methemoglobinemia (e.g., nitrites, dapsone, and benzocaine). The CO-oximetry is also indicated in these patients. Vascular dyes may also affect the S_pO_2 readings significantly, especially methylene blue, which is also used in the treatment of methemoglobinemia. Brown, blue, and green nail polish may affect S_pO_2 too. Therefore, routine removal of this polish is recommended. The issue of skin pigmentations effect on S_pO_2 reading is still controversial. Motion artifacts are a common problem in using pulse oximeters, especially in the intensive care units.

Future Directions for Pulse Oximeters

As mentioned above, there are several limitations with the recent commercially available pulse oximeters. Pulse oximeters technology is working on decreasing those limitations and improving pulse oximeters function (34). In the future, techniques to filter out the noise component common to both R and IR signals, such as Masimo signal extraction, will significantly decrease false alarm frequency. Pulse oximeters employing more than two wavelengths of light and more sophisticated algorithms will be able to detect dyshemoglobins. Improvements in reflection oximetry, which detects backscatter of light from light-emitting diodes placed adjacent to detectors, will allow the probes to be placed on any body site. Scanning of the retinal blood using reflection oximetry can be used as an index of cerebral oxygenation. Combinations of reflectance oximetry and laser Doppler flowmetry may be used to measure microcirculatory oxygenation and flow.

Continuous Intravascular Blood Gas Monitoring (CIBM)

The current standard for blood gas analysis is intermittent blood gas sampling, with measurements performed *in vitro* in the laboratory or by using bedside blood gas analyzer. Recently, miniaturized fiberoptic devices have been developed that can be placed intravascularly to continuously measure changes in PO_2, PCO_2, and pH. These devices utilize two different technologies: Electrochemical sensors technology, based on a modified Clark electrode, and optode (photochemical/optical) technology (35,36).

Optode (Photochemical–Optical) Technology. An optode unit consists of optical fibers with fluorescent dyes encased in a semipermeable membrane. Each analyte, such as hydrogen ion, oxygen, or carbon dioxide, crosses the membrane and equilibrates with a specific chemical fluorescent dye to form a complex. As the degree of fluorescence changes with the concentration of the analyte, the absorbance of a light signal sent through the fiberoptic bundles changes, and a different intensity light signal is returned to the microprocessor. Optode technology has accuracy comparable to that of a standard laboratory blood gas analyzer. However, several reasons and problems, including the cost (see below) still limit the use of this monitor routinely.

At present, the Paratrend 7+ (PT7+; Diametric Medical Inc., High Wycombe, U.K.; distributed by Philips Medical Systems), and Neotrend (NT) are the only commercially available multiparameter CIBM systems. The original probe of Paratrend 7 (PT7) was introduced in 1992. It consists of a hybrid probe incorporating four different sensors: miniaturized Clark electrode to measure PO_2, optode to determine PCO_2, and pH (absorbance sensors, phenol red in bicarbonate solution), and a thermocouple (copper, constantan) to measure temperature and allow temperature correction of the blood gas values. All these sensors were encased in a heparin-coated microporous polyethylene tube that was permeable to the analytes to be measured. This sensor was modified in 1999. In the new sensor (PT7+) (Fig. 20), the Clark electrode was replaced by an optical PO_2 sensor. According to the manufacturer, this new PO_2 sensor is more accurate and has a faster response time.

Clinical Uses

Continuous intravascular blood gas monitoring has been applied in various clinical settings (36,37) in the operating room and the intensive care unit. In the operating room, especially in adults undergoing one lung ventilation for major surgery (e.g., one lung ventilation for thoracoscopic surgery or lung transplantation, major cardiac or vascular surgery). The most common site for CIBM measurement is the radial artery in adults and the femoral artery in children. The umbilical artery is used for probe insertion in neonates. Reports and studies showed that performance and accuracy of CIBM devices appear to be sufficient for clinical use.

Figure 20. The Paratrend 7+ (PT7+; Diametric Medical Inc.) sensor.

Limitations and Complications

Reliable intravascular blood gas measurement depends on a number of mechanical, electrical, and physicochemical properties of the CIBM probe as well as the conditions of the vessel into which the probe is inserted (36,37). Therefore, several factors can affect the performance of CIBM, including mechanical factors related to the intraarterial probe (e.g., not advanced adequately in the artery, the sensor becomes attached to the wall of the vessel), factors related to the artery itself (e.g., vasospasm), interference from electrocautery and ambient or endoscopic light, or related to the "flush" solution used to flush the intraarterial catheter, which may lead to false measurements. Complications may include thrombosis, ischemia, vasospasm, and failure. Although CIBM appears to be advantageous, there are no prospective, randomized, double-blind studies of its impact on morbidity and mortality. Future outcome studies should focus on well-defined groups of selected patients who might benefit from CIBM (e.g., critically ill patients with potentially rapid and unexpected changes in blood gas values). Furthermore, no data is available on the cost/benefit ratio of CIBM, and more studies are still needed to know if this monitor is cost-effective.

Intrapartum Fetal Pulse Oximetry

Intrapartum fetal pulse oximetry is a direct continuous noninvasive method of monitoring fetal oxygenation (38). Persistent fetal hypoxemia may lead to acidosis and neurological injury, and current methods to confirm fetal compromise are indirect and nonspecific. Therefore, intrapartum fetal pulse oximetry may improve intrapartum fetal assessment and, most important, improve the specificity of detecting fetal compromise (39,40). Intrapartum fetal pulse oximetry may monitor, not only the fetal heart rate (FHR), but also the arterial oxygen saturation and peripheral perfusion may be assessed.

Principle of Operation and Placement

The fetus *in utero* does not have an exposed area that would allow placement of a transmission sensor (38). Thus, reflectance sensors have been designed where the light-emitting diodes are located adjacent to the photodetector (Fig. 9). During labor, the sensor is placed transvaginally between the uterine wall and the fetus, with contact on the fetal presenting part, usually the soft tissue of the fetal cheek. Monitoring of fetal oxygen saturation has been encumbered by multiple technical obstacles (38). For example, reflectance sensors not directly attached to the fetus, work only when in contact with fetal skin and may not produce an adequate S_pO_2 signal when contact is suboptimal during intense uterine contractions or during episodes of fetal movement. In this situation, sensor position may require adjustment. Improved reflectance sensor contact has been attempted via a variety of sensor modifications, including suction devices, application with glue, and direct attachment to the fetal skin with a special clip. The Nellcor (Fig. 21) sensors have been developed with a "fulcrum" modification, which mechanically places the sensor surface into better contact with the fetal skin. Other technical

Figure 21. Nellcor OxiFirst fetal pulse oximeter.

advances, such as modification of the red light-emitting diode from a 660 to a 735 nm wavelength, have resulted in improved registration times.

Future Direction of Intrapartum Fetal Pulse Oximetry. Ideally, calibration of these monitors in human fetuses should be done by simultaneous measurement of S_pO_2 and preductal S_aO_2. Because the access to fetal circulation during labor is not feasible, calibration of these monitors is still a major problem (38). It appears that well-designed animal laboratory studies and human infant and neonatal studies will have to suffice for calibration and validation of these monitors. To make this monitor more valuable and accurate as a guide for obstetric and neonatal management during labor, prospective studies with a larger number of abnormal fetuses will be necessary to determine duration and level of hypoxia leading to metabolic acidosis in humans. Also, more studies are needed to answer questions about its safety and efficacy. Finally, further refinements in equipment design should improve the accuracy of S_pO_2 determination and the ability to obtain an adequate signal. Decreased signal-to-noise ratios, motion artifacts (e.g., contractions, fetal movement, maternal movement), impediments to light transmission (e.g., vernix, fetal hair, meconium), and calibration difficulties are unique obstacles in accurately assessing the fetus by this monitor. Technical development goals of fetal pulse oximetry should include improvement of sensor optical design, hardware, and software modification to obtain high signal quality and precise calibration. Major advantages of fetal oxygen saturation monitoring include its ease of interpretation for clinicians of varying skills, being noninvasive method, and the ability to monitor fetal oxygenation continuously during labor. However, more studies are needed to evaluate its safety, efficacy, and cost issues (41). When these issues are resolved, intrapartum fetal oxygen saturation monitoring could perhaps be one of the major advances in obstetrics during the twenty-first century.

TRANSCUTANEOUS BLOOD GAS MONITORING (TCM)

Historical Development

The possibility of continuously monitoring arterial blood oxygen and carbon dioxide using a heated surface electrode on human skin was discovered in the early 1970s and made commercially available by 1976 (42). In 1951, Baumberger and Goodfriend published an article showing a method to determine the arterial oxygen tension in man by equilibration through intact skin. By immersing a finger in a phosphate buffer solution heated to 45 °C, they found that the PO_2 of the buffer approached that of the alveolar air. They showed that if skin blood flow increased by the highest tolerable heat (45°C), the surface PO_2 rises to arterial blood PO_2. A few years later (in 1956), Clark invented the membrane covered platinum polarographic electrode to measure O_2 tissue tensions. By 1977, at least three commercial transcutaneous PO_2 ($tcPO_2$) electrodes were available (Hellige, Roche, RADIOMETER). These devices were applied initially to premature infants in an effort to reduce the incidence of blindness due to excessive oxygen administration. Throughout more than three decades, the TCM technology has been closely linked to the care of neonates; however, recent studies suggest that TCM technology may work just as well for older children and adults (29). The TCM offers continuous noninvasive measurement of blood gases, which is especially advantageous in critically ill patients in whom rapid and frequently life-threatening cardiopulmonary changes can occur during short periods of time. However, with the widespread use of pulse oximetry, the use of transcutanenous blood gas monitors has decreased.

Blood Gas Diffusion Through the Skin

The human skin consists of three main layers: the stratum corneum, epidermis, and dermis (Fig. 22). The thickness of the human skin varies with age, sex, and region of the body. The thickness of the stratum corneum varies from 0.1 to 0.2 mm depending on the part of the body. This is nonliving layer composed mainly of dehydrated cells (dead layer), which do not consume oxygen or produce carbon dioxide. The next layer is the epidermis layer, which consists of proteins, lipids, and melanin-forming cells. The epidermis is living, but is blood-free. The thickness of this layer $\sim$0.05–1 mm. Underneath the epidermis is the dermis, which consists of dense connective tissue, hair follicles, sweat glands, fat cells, and capillaries. These capillaries receive blood from arterioles and drain in venules. Arteriovenous anastomoses innervated by nerve fibers are commonly found in the dermis of the palms, face, and ears. These shunting blood vessels regulate blood flow

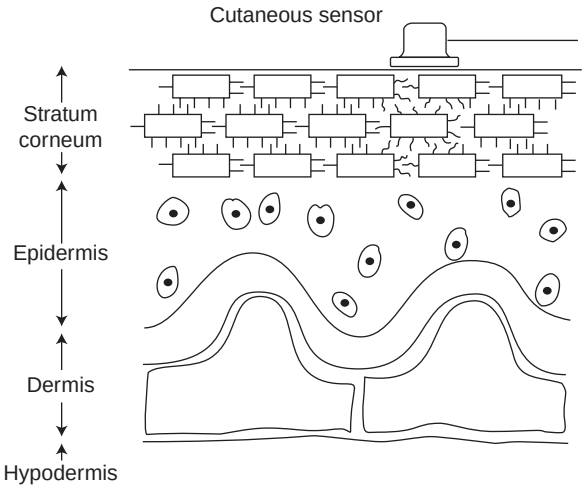


Figure 22. Human skin.

through the skin. Heat increases blood flow through these channels almost 30-fold. Gas diffusion through the skin occurs due to a partial pressure difference between the blood and the outermost surface of the skin. Diffusion of blood gases through the skin normally is very low, however, the heated skin ($\sim 43\,^{\circ}C$) becomes considerably more permeable to these gases.

Principle of Transcutaneous PO_2 Measurement (tcPO_2)

The probe used to measure the tcPO_2 is based on the idea of oxygen polarography (see above). This probe (7,12) consists of a platinum cathode and a silver reference anode encased in an electrolyte solution and separated from the skin by a membrane permeable to oxygen (usually made of Teflon, polypropylene, or polyethylene). The electrode is heated, thereby melting the crystalline structure of the stratum corneum, which otherwise makes this skin layer an effective barrier to oxygen diffusion. The heating of the skin also increases the blood flow in the capillaries underneath the electrodes. Oxygen diffuses from the capillary bed through the epidermis and the membrane into the probe, where it is reduced at the cathode, thereby generating an electric current that is converted into partial pressure measurements and displayed by the monitor. Because of an *in vitro* drift inevitably occurring inside the probe, where several chemical reactions are going on, the tcPO_2 sensor must be calibrated before using, and be repeated every 4–8 h. Since the O_2-dependent current flow exhibits a linear relationship at a fixed voltage, only two known gas mixtures are required for the calibration. Two *in vitro* calibration techniques can be employed: by using two precision gas mixtures (e.g., nitrogen and oxygen), and by using a "zero O_2 solution" (e.g., sodium sulfite) and room air.

The Transcutaneous PO_2 Sensor. The modern transcutaneous PO_2 sensors still use the principles used by Clark decades ago (7,12). Figure 23 illustrates a cross-sectional diagram of a typical Clark-type sensor. This particular sensor consists of three glass-sealed platinum cathodes that are separately connected via current amplifiers to a surrounding Ag–AgCl cylindrical ring. A buffered KCl electrode, which has a low water content to reduce drying of the sensor, is used. The following basic reactions happen between the two electrodes:

At the anode (+ electrode):

$$4\,Ag + 4\,Cl^- \leftrightarrow 4\,AgCl + 4\,e^-$$

(the electrons complete the circuit)

At the cathode (− electrode):

$$4\,H^+ + 4\,e^- + O_2 \leftrightarrow 2\,H_2O$$

(the electrons are boiled off of the platinum electrode)

Overall:

$$4\,Ag + 4\,Cl^- + 4\,H^+ + O_2 \leftrightarrow 4\,AgCl + 2\,H_2O$$

The two electrodes are covered with a thin layer of electrolytic solution that is maintained in place by a membrane that allows slow diffusion of O_2 from the skin into the sensor. The diffusion of O_2 through the skin is normally very low. Under normal physiological conditions, the PO_2 measured at the surface of the skin using a nonheated transcutaneous PO_2 electrode is near zero, regardless of the underlying blood PO_2. In order to facilitate O_2 diffusion through the skin, abrasion of the skin and drug-induced hyperemia through the application of nicotinic acid cream were initially used. However, since direct skin heating gives a more prolonged and consistent effect, a heating element is now used in all commercial transcutaneous PO_2 sensors. Generally, temperatures between 43 and 44° yield adequate vasodilatation of the cutaneous blood vessels with minimal skin damage. Heating the skin speeds up O_2 diffusion through the stratum corneum. In addition, it also causes vasodilatation of the dermal capillaries, which increases blood flow to the region of skin in contact with the sensor. With increased blood flow, more O_2 is available to the tissues surrounding the capillaries in the skin, and consequently the PO_2 of the blood in these capillary loops approximate more closely that of the arterial blood. Heating the blood also shifts the oxygen dissociation curve to the right. Therefore, the binding of hemoglobin with O_2 is reduced and the release of O_2 to the cells is increased. Simultaneously, skin heating also increases local tissue O_2 consumption. Fortunately, however, these two factors tend to cancel each other.

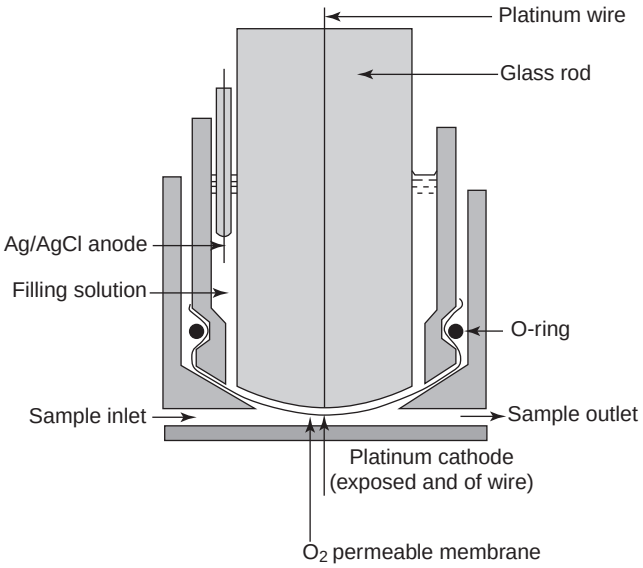


Figure 23. A cross-sectional diagram of a typical Clark-type sensor.

Transcutaneous PCO_2 Monitoring

Continuous PCO_2 monitoring is helpful in monitoring lung ventilation during spontaneous breathing or artificial ventilation. It makes it easier to adjust the parameters of the ventilator and prevent respiratory acidosis or alkalosis.

The Transcutaneous PCO_2 Sensor

The typical sensor is similar the O_2 sensor that was described above, as shown in Fig. 24. This sensor (7,12) consists of glass pH electrode with a concentric Ag–AgCl reference electrode that also serves as a temperature-

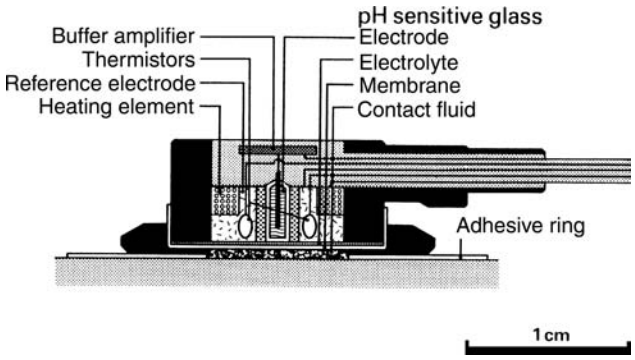


Figure 24. The transcutaneous PCO_2 sensor.

controlled heater. A buffer electrolyte (e.g., HCO_3^-) is placed on the surface of the electrode and a thin CO_2 permeable membrane (e.g., Teflon) stretched over the electrode separates the sensor from its surroundings. As CO_2 molecules diffuse via the CO_2-permeable membrane into the HCO_3^- containing solution, the following chemical reaction occurs:

$$CO_2 + H_2O \leftrightarrow H_2CO_3 \leftrightarrow H^+ + HCO_3^-$$

A potential between the pH and the reference electrodes is generated as a result of this reaction. This potential is proportional to the CO_2 concentration. Measurement of pH with a pH electrode can lead to estimation of the skin PCO_2, which correlates with the $PaCO_2$. According to the Henderson–Haselbach relationship, pH is proportional to the negative logarithm of PCO_2.

Skin temperature must be considered when analyzing PCO_2 measurements, because the skin heating can affect the transcutaneous PCO_2 sensor reading. This effect is due to the high temperature coefficient of the PCO_2 sensor. Heating the sensor results in an increase in PCO_2, since CO_2 solubility decreases, increase in local tissue metabolism, and increase in the rate of CO_2 diffusion through the stratum corneum. Therefore, the transcutaneous PCO_2 values are usually higher than the corresponding arterial PCO_2. Calibration of the PCO_2 sensor is different from the PO_2 sensor calibration. In the CO_2 sensor case, the voltage signal generated in the PCO_2 sensor is proportional to the logarithm of the CO_2 concentration (not to CO_2 concentration directly, as the case in PO_2). Therefore, there is no "zero point" calibration in transcutaneous PCO_2 sensor as there is with a transcutaneous PO_2 sensor. For this reason, one needs two different precisely analyzed gas mixtures for calibration. Usually, gas mixtures containing 5 and 10% CO_2 are used for calibrating the PCO_2 sensor. On the other side, PCO_2 sensor calibration must be done at the temperature at which it will be operated.

Clinical Applications of Transcutaneous PO_2 and PCO_2 Monitoring

Transcutaneous PO_2 and PCO_2 monitoring have found numerous applications in clinical medicine and research (42,43) during the past two decades: (1) neonatology: tcPO_2 monitoring remains the most commonly used technique to guide oxygen therapy in premature infants. In low birth weight infants, tcPO_2 is one of the best available monitor of ventilation. (2) Fetal monitoring: specially designed electrodes attached to the fetal scalp have been used. Changes in tcPO_2 rapidly reflected changing maternal and fetal conditions. Some studies showed that fetal tcPO_2 is considerably affected by local scalp blood flow, therefore repeated episodes of asphyxia, which may lead to increase in catecholamines, can reduce fetal scalp blood flow and lead to misleading reduction in tcPO_2. (3) Sleep studies: pulse oximetry and combined tcPO_2–tcPCO_2 electrode are used in sleep studies. This combination made it possible to study the ventilator response of hypoxia in sleeping infants. (4) Peripheral circulation: tcPO_2 electrodes are extensively used in evaluation peripheral vascular disease (44). Furthermore, transcutaneous oximetry has been used in several clinical situations such as prediction of healing potential for skin ulcers or amputation sites, assessment of microvascular disease (45), and determination of cutaneous vasomotor status. Figure 25 shows one of the commercially available transcutaneous blood gas monitor.

Figure 25. Radiometer TCM 4 transcutaneous blood gas monitor.

CAPNOMETRY AND CAPNOGRAPHY

Introduction

Capnometry is the measurement of carbon dioxide (CO_2) in the exhaled gas. Capnography is the method of displaying CO_2 measurements as waveforms (capnograms) during the respiratory cycle. The end-tidal PCO_2($PetCO_2$) is the maximum partial pressure of the exhaled CO_2 during tidal breathing (just before the beginning of inspiration). The measurement of CO_2 in respiratory gases was first accomplished in 1865, using the principle of Infrared(IR) absorption. Capnography was developed in 1943 and introduced to clinical practice in the 1950s (27). Since then, capnometry–capnography has gone through significant advances. Now capnography is a "standard of care" for general anesthesia (3), as described by the American Society of Anesthesiologists (ASA).

Measurement Techniques

Capnometry most commonly utilizes IR light absorption or mass spectrometry. Other technologies include Raman spectra analysis and a photoacoustic spectra technology (46,47)

Infrared Light Absorption Technique. This is the most common technique used to measure CO_2 in capnometers.

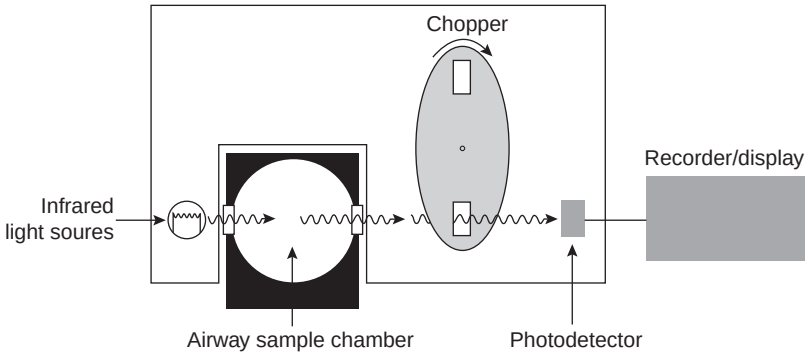


Figure 26. Single-beam infrared CO_2 analyzer, used in some mainstream sampling systems.

This method is cheaper and simpler than mass spectrometry. However, it is less accurate and has a slower response time ($\sim$0.25 vs. 0.1 s for mass spectrometry). There are two types of IR analyzers, a double and a single beam. The double-beam positive-filter model consists of an IR radiation source, which radiates to two mirrors. The two beams pass via a filter to two different chambers (sample chamber and reference chamber), and then to a photodetector. Consequently, it is possible to process the detector output electronically to indicate the concentration of CO_2 present. The single-beam negative-filter (Fig. 26), utilizes only one beam without using a reference. The principle behind this technique is that gases generally absorb electromagnetic IR radiation, and gas molecules with two or more atoms, provided these atoms are dissimilar (e.g., CO_2, but not O_2) absorb IR radiation in the range 1000–15000 nm. By filtering particular wavelengths, carbon dioxide and other gases can be measured. Carbon dioxide absorbs IR radiation strongly between 4200 and 4400 nm. Nitrous oxide and water have absorption peaks close to this area. Thus, there is a potential for the introduction of error with these substances in this method.

Raman Spectrography. Raman spectrography uses the principle of "Raman Scattering" for CO_2 measurement. The gas sample is aspirated into an analyzing chamber, where the sample is illuminated by a high intensity monochromatic argon laser beam. The light is absorbed by molecules, which are then excited to unstable vibrational or rotational energy states (Raman scattering). The Raman scattering signals (Raman light) are of low intensity and are measured at right angles to the laser beam. The spectrum of Raman scattering lines can be used to identify all types of molecules in the gas phase. Raman scattering technology has been incorporated into many newer anesthetic monitors (RASCAL monitors) to identify and quantify instantly CO_2 and inhalational agents used in anesthesia practice (48).

Mass Spectrography. The mass spectrograph separates molecules on the basis of mass to charge ratios. A gas sample is aspirated into a high vacuum chamber, where an electron beam ionizes and fragments the components of the sample. The ions are accelerated by an electric field into a final chamber, which has a magnetic field, perpendicular to the path of the ionized gas stream. In the magnetic field, the particles follow a path wherein the radius of curvature is proportional to the charge: mass ratio. A detector plate allows for determination of the components of the gas and for the concentration of each component. Mass spectrometers are quite expensive and too bulky to use at the bedside and are rarely used presently. They are either "stand alone", to monitor a single patient continuously, or "shared", to monitor gas samples sequentially from several patients in different locations (multiplexed). Up to 31 patients may be connected to a multiplexed system, and the gas is simultaneously sampled from all locations by a large vacuum pump. A rotary valve (multiplexer) is used to direct the gas samples sequentially to the mass spectrometer. In a typical 16-station system, with an average breathing rate of 10 breaths $\cdot$ min^{-1}, each patient will be monitored about every 3.2 min. The user can interrupt the normal sequence of the multiplexer and call the mass spectrometer to his patient for a brief period of time (46–48).

Photoacoustic Spectrography. Photoacoustic gas measurement is based on the same principles as conventional IR-based gas analyzers: the ability of CO_2, N_2O and anesthetic agents to absorb IR light (46,49). However, they differ in measurement techniques. While IR spectrography uses optical methods, photoacoustic spectrography (PAS) uses an acoustic technique. When an IR energy is applied to a gas, the gas will expand and lead to an increase in pressure. If the applied energy is delivered in pulses, the gas expansion would be also pulsatile, resulting in pressure fluctuations. If the pulsation frequency lies within the audible range, an acoustic signal is produced and is detected by a microphone. Potential advantages of PAS over IR spectrometry are higher accuracy, better reliability, less need of preventive maintenance, and less frequent need for calibration. Furthermore, as PAS directly measures the amount of IR light absorbed, no reference cell is needed and zero drift is nonexistent in PAS. The zero is reached when there is no gas present in the chamber. If no gas is present there can be no acoustic signal (49).

CO_2 Sampling Techniques

Sidestream versus Mainstream. Capnometers that are used in clinical practice use two different sampling techniques (50) (Fig. 27): sidestream or mainstream. A mainstream (flow-through) capnometer has an airway adaptor cuvette attached in-line and close to the endotracheal tube. The cuvette incorporates an IR light source and sensor that

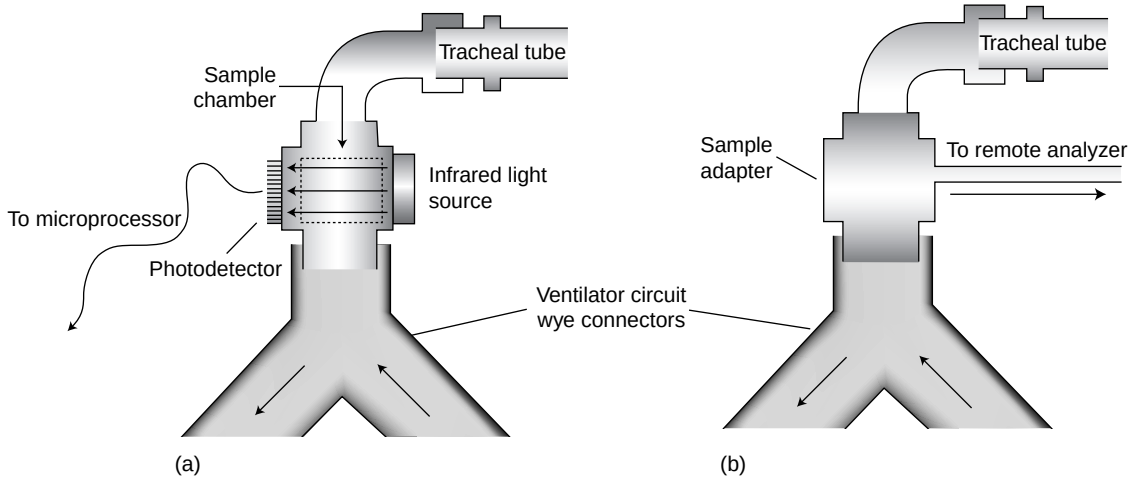


Figure 27. Sidestream vs. mainstream CO_2 sampling techniques. (a) Mainstream CO_2 sampling. (b) Sidestream sampling.

senses carbon dioxide absorption to measure PetCO_2. A sidestream capnometer uses a sampling line that attaches to a T-piece adapter at the airway opening, through which the instrument continually aspirates tidal airway gas for analysis of carbon dioxide. The main advantage of the mainstream analyzer is its rapid response, because the measurement chamber is part of the breathing circuit. The sample cuvette lumen, through which inspired and expired gases pass, is large in order to minimize the work of breathing, and pulmonary secretions generally do not interfere with carbon dioxide analysis. Compared with sidestream (aspiration) sampling, the airway cuvette is relatively bulky and can add dead space. However, within the past few years lighter and smaller airway cuvettes have been developed to allow its use in neonates. The sidestream PCO_2 analyzer adds only a light T-adapter to the breathing circuit, and can be easily adapted to non-intubation forms of airway control. Because the sampling tubing is small bore, it can be blocked by secretions. During sidestream capnography, the dynamic response, the steepness of the expiratory upstroke and aspiratory downslope, tends to be blunted because of the dispersive mixing of gases through the sampling line, where gas of high PCO_2 mixes with gas of low PCO_2. In addition, a washout time is required for the incoming sampled gas to flush out the volume of the measuring chamber. The overall effect is an averaging of the capnogram, resulting in a lowering of the alveolar plateau and an elevation of the inspiratory baseline. Thus, PetCO_2 may be underestimated and rebreathing can be simulated. These problems are exacerbated by high ventilatory rates and by the use of long sampling catheters. In addition, the capnogram is delayed in time by transport delay, the time required to aspirate gas from the airway opening adapter through the sampling tubing to the sampling chamber.

Micro-Stream Technology. Micro-stream technology (51) is a new CO_2 sampling techneque that uses a low aspiration rate (as low as 50 $mL \cdot min^{-1}$), such as NBP-75, Nellcor Puritan Bennett, as shown in Fig. 28. In addition, this technology uses a highly CO_2-specific IR source, where the IR emission exactly matches the absorption spectrum of the CO_2 molecules. The advantages of this technology, compared to the traditional high flow side-stream capnometer (150 $mL \cdot min^{-1}$), is that it gives more accurate PetCO_2 measurements and better waveforms in neonates and infants with small tidal volumes and high respiratory rates. Furthermore, these low flow capnometers are less likely to aspirate water and secretions into the sampling tubes, resulting in either erroneous PetCO_2 values or in total occlusion of sampling tube.

Phases of Capnography

A normal single breath capnogram (time capnogram) is shown in Fig. 29. Time capnogram is the partial pressure of expired CO_2 plotted against time on the horizontal axis. This capnogram can be divided into inspiratory (phase 0) and expiratory segments. The expiratory segment, similar to a single breath nitrogen curve or single breath CO_2 curve, is divided into phases I, II, and III, and occasionally, phase IV, which represents the terminal rise in CO_2 concentration. The angle between phase II and III is the alpha angle. The nearly 90° angle between phase III and the descending limb is the beta angle. Changes in time

Figure 28. Nellcor Microstream $ETCO_2$ breath sampling unit.

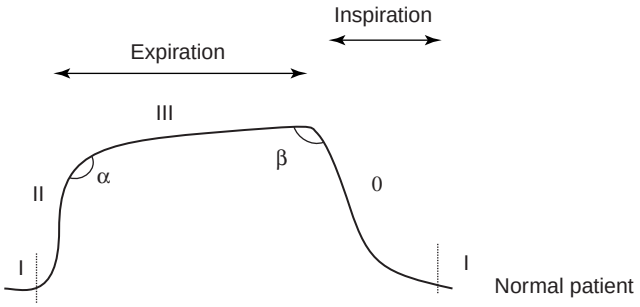


Figure 29. Normal single breath capnogram.

capnogram help to diagnose some of the breathing and ventilation problems, especially during anesthetic management of patients undergoing surgery (e.g., bronchospasm, esophageal intubation, CO_2 rebreathing, and even cardiac arrest).

Clinical Uses of Capnography

Capnography and capnometry (52) are safe, noninvasive test, and have few hazards. They are widely used in clinical medicine. Their uses include, but are not limited to (*1*) evaluating the exhaled CO_2, especially end-tidal CO_2 in mechanically ventilated patients during anesthesia. (*2*) Monitoring the severity of pulmonary disease and evaluating response to therapy, especially therapy intended to improve the ratio of dead space to tidal volume (V_D/V_T) and the matching of ventilation to perfusion (V/Q). (*3*) Determining that tracheal rather than esophageal intubation has taken place (low or absent cardiac output may negate its use for this indication) (53). Colorimetric CO_2 detectors are adequate devices for this purpose. (*4*) Evaluating the efficiency of mechanical ventilatory support by determination of the difference between the arterial partial pressure for CO_2 (PCO_2) and the $PetCO_2$. Figure 30 shows a combined handheld capnograph/pulse oximeter.

Limitations

Note that although the capnograph provides valuable information (52) about the efficiency of ventilation, it is not a replacement or substitute for assessing the PCO_2. The difference between PCO_2 and $PetCO_2$ increases as dead-space volume increases. In fact, the difference between the PCO_2 and $PetCO_2$ has been shown to vary within the same patient over time. Alterations in breathing pattern and tidal volume may introduce error into measurements designed to be made during stable, steady-state conditions. Interpretation of results must take into account the stability of physiologic parameters, such as minute ventilation, tidal volume, cardiac output, ventilation/perfusion ratios, and CO_2 body stores. Certain situations may affect the reliability of the capnogram. The extent to which the reliability is affected varies somewhat among types of devices (IR, photoacoustic, mass spectrometry, and Raman spectrometry). Furthermore, the composition of the respiratory gas mixture may affect the capnogram (depending on

Figure 30. Nellcor OxiMax NBP-75 handheld capnograph/ pulse oximeter.

the measurement technology incorporated). The IR spectrum of CO_2 has some similarities to the spectra for both oxygen and nitrous oxide. High concentrations of either or both oxygen or nitrous oxide may affect the capnogram, and, therefore, a correction factor should be incorporated into the calibration of any capnograph used in such a setting. The reporting algorithm of some devices (primarily mass spectrometers) assumes that the only gases present in the sample are those that the device is capable of measuring. When a gas that the mass spectrometer cannot detect (such as helium) is present, the reported values of CO_2 are incorrectly elevated in proportion to the concentration of helium in the gas mixture. Moreover, the breathing frequency may affect the capnograph. High breathing frequencies may exceed the response capabilities of the capnograph. In addition, the breathing frequency, > 10 breaths $\cdot$ min^{-1}, has been shown to affect devices differently. Contamination of the monitor or sampling system by secretions or condensate, a sample tube of excessive length, a sampling rate that is too high, or obstruction of the sampling chamber, can lead to unreliable results. Use of filters between the patient airway and the sampling line of the capnograph may lead to lowered $PetCO_2$ readings. Inaccurate measurement of expired CO_2 may be caused by leaks of gas from the patient–ventilator system preventing collection of expired gases, including, leaks in the ventilator circuit, leaks around tracheal tube cuffs, or uncuffed tracheal tubes.

Sublingual Capnometry

Sublingual capnometry is a method to measure the partial pressure of carbon dioxide under the tongue ($PSLCO_2$). This method is being used mainly in the critical care units to evaluate patients with poor tissue perfusion and multiple organ dysfunction syndrome.

Pathophysiologic Basis. Significant increases in the partial pressure of carbon dioxide (PCO_2) in tissue have

Figure 31. Nellcor CapnoProbe Sublingual capnometer.

been associated with hypoperfusion, tissue hypoxia, and multiple organ dysfunction syndrome (54). When perfusion of the intestinal mucosa is compromised, CO_2 accumulates in the gut. The high diffusability of CO_2 allows for rapid equilibration of PCO_2 throughout the entire gastrointestinal (GI) tract. The vasculature of the tongue and the GI tract are controlled by similar neuronal pathways. Thus, the vasculatures of both respond similarly during vasoconstriction (55). Because the tongue is the most proximal part of the GI tract, measurement of PCO_2 can be conveniently and noninvasively obtained by placing a sensor under the tongue. Clinical studies have demonstrated that $PSLCO_2$ can be used in the assessment of systemic tissue hypoperfusion and hypercapnia (56).

Capnometer Components and Principle of Operation. Figure 31 shows a commercially available sublingual capnometer (Nellcor CapnoProbe Sublingual System). This system (25) consists of two components: (1) SLS-l Sublingual Sensor: This sensor contains an optrode (a sensitive analyte detector) consists of an optical fiber capped with a small silicone membrane containing a pH-sensitive solution (Fig. 32). When the optrode is brought into contact with sublingual tissue, CO_2 present in the tissue freely diffuses across the silicone membrane into the fluorescent dye solution. No other commonly encountered gases or liquids can pass across the membrane. The CO_2 dissolves and forms carbonic acid, which in turn lowers the pH of the solution. The fluorescence intensity of the dye in the solution is directly proportional to pH.

This single use sensor is packaged in a sealed metal canister. Inside the canister, the sensor tip is enclosed in a gas permeable reservoir that contains a buffer solution. The solution prevents the optrode from drying out. The solution also allows calibration just prior to use, as it is in equilibrium with a known concentration of CO_2 within the canister. To begin use, the clinician opens the canister and inserts the cable handle into the SLS-l Sublingual Sensor. This action initiates a calibration cycle that allows the

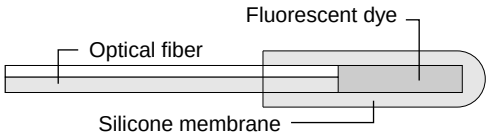


Figure 32. Diagram of Nellcor CapnoProbe sublingual capnometer basic components.

instrument to observe the sensor signal at the known PCO_2 of the calibrant (2). The N-80 instrument contains a precision optical component that emits light at two wavelengths in the violet and blue portions of the visible spectrum. The fluorescence intensity generated by the violet wavelength is insensitive to pH, whereas that generated by the blue wavelength is strongly sensitive to pH. The light is launched into an optical fiber and delivered to the tip of the disposable sensor. The green fluorescent light generated in the optrode is directed back to the N-80 instrument through an optical fiber. The light is then ratio metrically quantitated and directly correlated to $PSLCO_2$. When the fluorescence intensity of the optrode has stabilized (within 60–90 s), the N-80 instrument reports the measured value of $PSLCO_2$ on its LCD screen.

Chemical Colorimetric Airway Detector

This is a device (57,58) that uses a pH-sensitive indicator to detect breath-by-breath exhaled carbon dioxide (Fig. 33).

Figure 33. Nellcor Easy Cap II Pedi-Cap chemical colorimetric CO_2 detector.

The colorimetric airway detector is interposed between the endotracheal tube (ETT) and the ventilation device. Both adult and pediatric adaptors exist, but they cannot be used in infants who weigh < 1 kg. Because of excessive flow resistance, they are not suited for patients who are able to breathe spontaneously. Excessive humidity will render them inoperative in 15–20 min. The devices can be damaged by mucous, edematous, or gastric contents, and by administration of intratracheal epinephrine. Despite these drawbacks, colorimetric sensors have been found to be useful in guiding prehospital CPR (cardiopulmonary resuscitation) both in intubated patients and those with a laryngeal mask airway.

Role of Capnometry–Capnography in CPR. The relationship between cardiac output and $P\text{et}CO_2$ is logarithmic (59). Decreased presentation of CO_2 to the lungs is the major rate-limiting determinant of the $P\text{et}CO_2$ during low pulmonary blood flow. Capnography can detect the presence of pulmonary blood flow even in the absence of major pulses (pseudoelectromechanical dissociation, EMD) and also can rapidly indicate changes in pulmonary blood flow (cardiac output) caused be alterations in cardiac rhythm. Data suggests that $P\text{et}CO_2$ correlates with coronary perfusion pressure, cerebral perfusion pressure, and blood flow during CPR. This correlation between perfusion pressure and $P\text{et}CO_2$ is likely to be secondary to the relationship of $P\text{et}CO_2$ and cardiac output (60).

SUMMARY

Blood gas measurement methods and instruments have gone through significant improvements and advances in the last few decades. Invasive techniques have moved steadily toward using smaller instruments and closer to the patient's bed (bedside), which requires smaller blood sample. These improvements have made these devices more convenient, need less personnel to operate them, and more cost-effective. These bedside devices have comparable accuracy and reliability to the traditional central laboratory instruments. Continuous intravascular blood gas monitoring is a new invasive technique that uses miniaturized fiberoptic devices. This method has been used in different clinical settings with good results. However, several limitations and complications still exist. More studies and improvements are needed to know its cost-effectiveness in clinical medicine and to minimize its complications (such as ischemia and thrombosis). Other invasive instruments and methods were discussed in other parts of this encyclopedia. On the other hand, noninvasive blood gas measurement methods and devices improved greatly since its instruction to clinical medicine. These devices have been used extensively during anesthesia administration and in critical care units. Using pulse oximetry is "standard of care" in anesthesia practice. The use of pulse oximeter decreased the risk of hypoxia and its deleterious effect significantly. However, several limitations to its use still exist, especially in low perfusion states, during the presence of dyshemoglobins and motion artifacts. New technologies, such as the Oxi-Max and the Signal Extraction technologies, have been developed to overcome some of these limitations, and to add more features for these instruments (such as scanning the retina as an index of cerebral oxygenation and using Laser Doppler flowmetry–reflection oximetry to measure microcirculatory oxygenation and flow). Other types of pulse oximetry have been introduced to clinical medicine. Intrapartum fetal pulse oximetry is an example of these new pulse oximeters. This device provides a continuous, noninvasive method of monitoring fetal oxygenation, which may help in detecting persistent fetal hypoxemia and improve intrapartum fetal assessment. However, more studies are needed to evaluate its safety, efficacy, and cost issues. Transcutaneous blood gas monitoring is another noninvasive method to measure blood gases. This method is losing ground and popularity against the newer pulse oximeters, which have replaced this method in several situations. Capnometry–capnography has been used extensively in anesthesia practice in the last two decades. New CO_2 sampling technique such as microstream technology has been introduced. This method uses a low aspiration rate, making it more accurate than the previous techniques. Moreover, this technique can detect very small amount of CO_2. Photoacoustic sepectrography is a new technique has been developed to measure $P\text{et}CO_2$. This method is more reliable, accurate, and needs less calibration than the traditional methods. Studies showed an encouraging result and an important role for capnometry–capnography in cardiopulmonary resuscitation, which may lead to widespread use of these devices in CPR, in prehospital and inhospital settings.

BIBLIOGRAPHY

Cited References

1. Cadden K, Norman E, Booth J. Use of ABG in trauma for early recognition of acidosis and hypoxemia (Abstract). Respir Care 2001;46:1106.
2. Levin KP, Hanusa BH, Rotondi A, Singer DE, et al. Arterial blood gas and pulse oximetry in initial management of patients with community-acquired pneumonia. J Gen Intern Med 2001;9:590.
3. Morgan Jr GE, Mikhail MS, Murray MJ. Clinical Anesthesiology. 3rd ed. New York: Lange Medical Books/McGraw-Hill; 2002. p 124–125.
4. Severinghaus JW, Astrup P, Murray JF. Blood gas analysis and critical care medicine. Am J Resp Crit Care Med 1998;157:S114–S122.
5. Schell RM, Cole DJ. Cerebral monitoring: Jugular venous oximetry. Anesth Analg 2000;90:559.
6. Goodwin ALP. Physics of Gases, Anaesthesia and Intensive Care Medicine. (UK): The Medicine Publishing Company, Ltd.; 2003.
7. Malley WJ. Clinical Blood Gases: Assessment and Intervention. 2nd ed. New York: Elsevier; 2005.
8. West JB. Respiratory Physiology-Essentials. 6th ed. Baltimore: Williams & Wilkins; 2000.
9. Murray JF, Nadel JA. Textbook of Respiratory Medicine. 3rd ed. New York: W. B. Saunders.
10. Severinghaus JW, Astrup PB. History of blood gas analysis. VI: Oximetry. J Clin Monit 1986;2:270–288.
11. Millikan GA, Pappenheimer JR, Rawson AJ, et al. Continuous measurement of oxygen saturation in man. Am J Physiol 1941;133:390.

12. Adams AP, Hahn CEW. Principles and Practice of Blood Gas Analysis. London: Franklin Scientific Products; 1979.
13. Payne JB, Severinghaus JW. Pulse Oximetry. New York: Springer-Verlag; 1986.
14. Baker SJ, Tremper KK. The effect of carbon monoxide inhalation on pulse oximetry and transcutaneous PO_2. Anesthesiology 1987;66:677–679.
15. Baker SJ, Tremper KK, Hyatt J. Effects of methemoglobinemia on pulse oximetry and mixed venous oximetry. Anesthesiology 1989;70:112–117.
16. Severinghaus JW, Honda Y. History of blood gas analysis. VII. Pulse oximetry. J Clin Monit 1987;3:135–138.
17. Severinghaus JW. History and recent developments in pulse oximetry. Scand J Clin Lab Invest 1993;214 (1 Suppl): 105–111.
18. Merrick EB, Hayes TJ. Continuous noninvasive measurements of arterial blood oxygen levels. Hewlett-Packard J 1976;28920:2–9.
19. Aoyagi T, Miyasaka K. Pulse oximetry: its invention, contribution to medicine, and future tasks. Anesth Analg 2002;94(1 Suppl): S1–3.
20. Tremper KK, Barker SJ. Pulse oximetry. Anesthesiology 1989;70:98–108.
21. Welch JP, DeCesare MS, Hess D. Pulse oximetry: Instrumentation and clinical applications. Respir Care 1990;35: 584–601.
22. Robertson F, Hoffman G. Clinical evaluation of Masimo SET and Nellcor N395 oximeters during signal conditions in difficult-to-monitor neonates. Anesthesiology 2002;96:A556.
23. Barker SJ. The performance of six "motion-resistant" pulse oximeters during motion, hypoxemia, and low perfusion in volunteers. Anesthesiology 2001;95:A587.
24. Pologe JA, Tobin RM. Method and apparatus for improved photoplethysmographic perfusion-index monitoring. US patent 5,766,127. 1998.
25. http://www.nellcor.com.
26. Lima AP, Beelen P, Bakker J. Use of peripheral perfusion index derived from the pulse oximetry signal as a noninvasive indicator of perfusion. Crit Care Med 2002;30(6):1210–1213.
27. Soubani AO. Noninvasive monitoring of oxygen and carbon dioxide. Am J Emerg Med 2001;19(2).
28. Shapiro BA, Harrison RA, Cane RD. Clinical Application of Blood Gases. 4th ed. Chicago: Year Book Medical Publishers, Inc.; 1989.
29. Williams AJ. ABC of oxygen: Assessing and interpreting arterial blood gases and acid–base balance. Br Med J 1998;317(7167): 1213–1216.
30. Poets FC, Southall DP. Noninvasive monitoring of oxygenation in infants and children: Practical considerations and areas of concern. Pediatrics 1994;93(5): 737–746.
31. Severinghaus JW, Naifeh KH, Koh SO. Errors in 14 pulse oximeters during profound hypoxia. J Clin Monit 1989;5: 72–81.
32. Clayton DG, Webb RK, Ralston AC, et al. A comparison of the performance of 20 pulse oximeters under conditions of poor perfusion. Anaesthesia 1991;46:3–10.
33. Sidi A, Paulus DA, Rush W, et al. Methylene blue and indocyanine green artifactually low pulse oximetry readings of oxygen saturation. Studies in dogs. J Clin Monit 1987; 3:249–256.
34. Lynn LA. Interpretive Oximetry: Future Directions for Diagnostic Applications of SpO_2 Time-Series. Anesthesia Analgesia 2002;94:S84–S88.
35. Zimmerman JL, Dellinger RP. Initial evaluation of a new intra-arterial blood gas system in humans. Crit Care Med 1993;21:495–500.
36. Ganter M, Zollinger A. Continuous intravascular blood gas monitoring: development, current techniques, and clinical use of a commercial device. Br J Anaesth 2003;91:397–407.
37. Coule LW, Truemper EJ, Steinhart CM, Lutin WA. Accuracy and utility of a continuous intra-arterial blood gas monitoring system in pediatric patients. Crit Care Med 2001;29(2): 420–426.
38. Dildy GA. Intrapartum fetal pulse oximetry: past, present, and future. Am J Obstet Gynecol 1996;175(1): 1–9.
39. Dildy GA, Van den Berg PP, Katz M, et al. Intrapartum fetal pulse oximetry: fetal oxygen saturation trends during labor and relation to delivery outcome. Am J Obstet Gynecol 1994; 171:679–684.
40. Papiernik E. Fetal pulse oximetry: correlation between changes in oxygen saturation and neonatal outcome. Eur J Obstet Gynecol Reprod Biol 1994;57:73–77.
41. McNamara H, Chung DC, Lilford R, Johnson N. Do fetal pulse oximetry readings at delivery correlate with cord blood oxygenation and acidaemia. Br J Obstet Gynaecol 1992;99: 735–738.
42. Severinghaus JW. The current status of transcutaneous blood gas analysis and monitoring. Blood Gas News 1998; 9(2).
43. Franklin ML. Transcutaneous measurement of partial pressure of oxygen and carbon dioxide. Respir Care Clin North Am 1995;1:119–131.
44. Padberg FT, Back TL, Thompson PN, et al. Transcutaneous oxygen ($TcPO_2$) estimates probability of healing in the ischemic extremity. J Surg Res 1996;60:365–369.
45. Rooke TW. The use of transcutaneous oximetry in the noninvasive vascular laboratory. Int Angiol 1992;11(1): 46–40.
46. Tremper KK, Barker SJ. Fundamental principles of monitoring instrumentation. In: Miller RD, editor. Anesthesia. Vol I, 3rd ed. New York: Churchill Livingstone; 1990. p 957–999.
47. Raemer BD, Philip JH. Monitoring anesthetic and respiratory gases. In: Blitt CE, editor. Monitoring in Anesthesia and Critical Care Medicine. 2nd ed. New York: Churchill Livingstone; 1990. p 373–386.
48. Graybeal JM, Russell GB. Relative agreement between Raman and mass spectrometry for measuring end-tidal carbon dioxide. Respir Care 1994;39:190–194.
49. Mollgaard K. Acoustic gas measurement. Biomed Instr Technol 1989;23:495–497.
50. Block FE, McDonald JS. Sidestream versus mainstream carbon dioxide analyzers. J Clin Monit 1992;8:139–141.
51. Casti A, Gallioli G, Scandroglio G, Passaretta R, Borghi B, Torri G. Accuracy of end-tidal carbon dioxide monitoring using the NBP-75 microstream capnometer. A study in intubated ventilated and spontaneously breathing nonintubated patients. Euro J Anesthesiol 2000;17:622–626.
52. AARC Clinical Practice Guidelines. Capnography/capnometry during mechanical ventilation (2003 update). Respir Care 2003;48:534–539.
53. Shibutani K, Muraoka M, Shirasaki S, Kubal K, Sanchala VT, Gupte P. Do changes in end-tidal PCO_2 quantitatively reflect changes in cardiac output? Anesth Analg 1994;79(5): 829–833.
54. Rackow EC, et al. Sublingual capnometry and indexes of tissue perfusion in patients with circulatory failure. Chest 2001;120:1633–1638.
55. Weil MB, et al. Sublingual capnometry: a new noninvasive measurement for diagnosis and quantitation of severity of circulatory shock. Crit Care Med 1999;27:1225–1229.
56. Marik PE. Sublingual capnography: a clinical validation study. Chest 2001;120:923–927.

57. Kelly JS, Wilhoit RD, Brown RE, James R. Efficacy of the FEF colorimetric end-tidal carbon dioxide detector in children. Anesth Analg 1992;75:45–50.
58. Nakatani K, Yukioka H, Fujimori M, et al. Utility of colorimetric end-tidal carbon dioxide detector for monitoring during prehospital cardiopulmonary resuscitation. Am J Emerg Med 1999;17:203–206.
59. Ornato JP, Garnett AR, Glauser FL, Virginia R. Relationship between cardiac output and the end-tidal carbondioxide tension. Ann Emerg Med 1990;19:1104–1106.
60. White RD, Asplin BR. Out of hospital quantitative monitoring of end-tidal carbondioxide pressure during CPR. Ann Emerg Med 1994;23:25–30.

See also CHROMATOGRAPHY; FIBER OPTICS IN MEDICINE; PERIPHERAL VASCULAR NONINVASIVE MEASUREMENTS.

BLOOD PRESSURE MEASUREMENT

CAN ISIK
Electrical Engineering and Computer Science Department, Syracuse University Syracuse, New York

INTRODUCTION

Blood pressure is an important signal in determining the functional integrity of the cardiovascular system. Scientists and physicians have been interested in blood pressure measurement for a long time. The first blood pressure measurement is attributed to Reverend Stephen Hales, who in the early eighteenth century connected water-filled glass tubes in the arteries of animals and correlated their blood pressures to the height of the column of fluid in the tubes. It was not until the early twentieth century that the blood pressure measurement was introduced into clinical medicine, albeit with many limitations.

Blood pressure measurement techniques are generally put into two broad classes: direct and indirect. Direct techniques of blood pressure measurement, which are also known as invasive techniques, involve a catheter to be inserted into the vascular system. The indirect techniques are noninvasive, with improved patient comfort and safety, but at the expense of accuracy. The accuracy gap between the invasive and the noninvasive methods, however, has been narrowing with the increasing computational power available in portable units, which can crunch elaborate signal processing algorithms in a fraction of a second.

During a cardiac cycle, blood pressure goes through changes, which correspond to the contraction and relaxation of the cardiac muscle, with terminology that identifies different aspects of the cycle. The maximum and minimum pressures over a cardiac cycle are called the systolic and diastolic pressures, respectively. The time average of the cardiac pressure over a cycle is called the mean pressure, and the difference between the systolic and diastolic pressures is called the pulse pressure.

Normal blood pressure varies with age, state of health, and other individual conditions. An infant's typical blood pressure is 80/50 mmHg (10.66/6.66 kPa) (systolic/diastolic). The normal blood pressure increases gradually and reaches 120/80 (15.99/10.66 kPa) for a young adult. Blood pressure is lower during sleep and during pregnancy. Many people experience higher blood pressures in the medical clinic, a phenomenon called the "white coat effect." Therefore, the ranges given in Table 1 are used as guidelines rather than as diagnostic facts.

Table 1. Classification of Blood Pressure for Adults

Category	Systolic—mmHg		Diastolic—mmHg
Normal	<120	and	<80
Prehypertension	120–139	or	80–89
Stage 1 Hypertension	140–159	or	90–99
Stage 2 Hypertension	160 or higher	or	100 or higher

DIRECT TECHNIQUES

The operation of direct measurement techniques can be summarized in very simple terms: They all use a pressure transducer that is coupled to the vascular system through a catheter or cannula that is inserted to a blood vessel, followed by a microcontroller unit with electronics and algorithms for signal conditioning, signal processing, and decision making. There are many advantages of this set of techniques, including:

- The pressure is measured very rapidly, usually within one cardiac cycle.
- The measurement is done to a very high level of accuracy and repeatability.
- The measurement is continuous, resulting in a graph of pressure against time.
- The measurement is motion tolerant.

Therefore, the direct techniques are used when it is necessary to accurately monitor patients' vital signs, for example, during critical care and in the operating room. Although direct techniques have a lot in common, there are differences in the details of various approaches.

Extravascular Transducers

The catheter in this type of device is filled with a saline solution, which transmits the pressure to a chamber that houses the transducer assembly. As a minor disadvantage, this structure affects the measured pressure through the dynamic behavior of the catheter. As the catheter has a known behavior, this effect can be minimized to insignificant levels through computational compensation (1).

Intravascular Transducers

The transducer is at the tip of the catheter in this type of device. Then the measured signal is not affected by the hydraulics of the fluid in the catheter. The catheter diameter is larger in this class of transducers.

Transducer Technology

A wide spectrum of transducer technologies is available to build either kind of transducer. They include metallic or semiconductor strain is gauges, piezoelectric, variable capacitance, variable inductance, and optical fibers. Appropriate driver and interface circuitry accompanies each technology (2).

Other Applications of Direct Pressure Measurement

Another advantage of direct measurement techniques is that they are not limited to measuring the simple arterial pressure. They can be used to obtain central venous, pulmonary arterial, left atrial, right atrial, femoral arterial, umbilical venous, umbilical arterial, and intracranial pressures by inserting the catheter in the desired site (3).

Sources of Errors

Direct blood pressure measurement systems have the flexibility of working with a variety of transducers/probes. It is important that the probes are matched with the appropriate compensation algorithm. Most modern equipment does this matching automatically, eliminating the possibility of operator error. An additional source of error occurs when air bubbles get trapped in the catheter. This changes the fluid dynamics of the catheter, causing an unintended mismatch between the catheter and its signal processing algorithm. This may cause distortions in the waveforms and errors in the numeric pressure values extracted from them. It is difficult to recognize this artifact from the waveforms, so it is best to avoid air bubbles in the catheter.

NONINVASIVE (INDIRECT) TECHNIQUES

An overwhelming majority of blood pressure measurements do not require continuous monitoring or extreme accuracy. Therefore, noninvasive techniques are used in most cases, maximizing patient comfort and safety. Currently available devices for noninvasive measurement are

- Manual devices: These devices use the auscultatory technique.
- Semiautomatic devices: These devices use oscillatory techniques.
- Automatic devices: Although most of these devices use oscillatory techniques, some use pulse-wave velocity or plethysmographic methods.

The Auscultatory Technique

In the traditional, manual, indirect measurement system, an occluding cuff is inflated and a stethoscope is used to listen to the sounds made by the blood flow in the arteries, called Korotkov sounds. When the cuff pressure is above the systolic pressure, blood cannot flow, and no sound is heard. When the cuff pressure is below the diastolic pressure, again, no sound is heard. A manometer connected to the cuff is used to identify the pressures where the transitions from silence to sound to silence are made. This combination of a cuff, an inflating bulb with a release valve, and a manometer is called a sphygmomanometer and the method an auscultatory technique. Usually, the cuff is placed right above the elbow, elevated to the approximate height of the heart, and the stethoscope is placed over the brachial artery. It is possible to palpate the presence of pulse under the cuff, rather than to use a stethoscope to listen to the sounds. The latter approach works especially well in noisy places where it is hard to hear the heart sounds.

Figure 1. Blood pressure waveform, and systolic, diastolic, and mean pressures, from an invasive monitor screen (4).

This method has various sources of potential error. Most of these sources are due to misplacement of the cuff, problems with hearing soft sounds, and using the wrong cuff size. Using a small cuff on a large size arm would result in overestimating the blood pressure, and vice versa. Nevertheless, an auscultatory measurement performed by an expert healthcare professional using a clinical grade sphygmomanometer is considered to be the gold standard in noninvasive measurements.

Oscillatory Techniques

Most automatic devices base their blood pressure estimations on the variations in the pressure of the occluding cuff, as the cuff is inflated or deflated. These variations are due to the combination of two effects: the controlled inflation or deflation of the cuff and the effect of the arterial pressure changes under the cuff. The Korotkov sounds are not used in the oscillatory techniques.

The cuff pressure variation data may be collected while the cuff is being inflated or deflated. Furthermore, the inflation or deflation during the data collection may be controlled in a continuous fashion or in a step-wise fashion. This variability gives four different strategies in data collection. Their differences may seem insignificant at first, but they have significant effects on the way a variety of algorithms are designed.

Data in Fig. 2 were collected using an experimental system. The cuff is first rapidly inflated to a value higher than the anticipated systolic pressure, an approximate pressure of 170 mmHg (22.66 kPa) in this case. Then it is deflated in small steps until the cuff pressure is below the anticipated diastolic pressure, $\sim$50 mmHg (6.66 kPa). Please note that when the cuff pressure is very high or very low, the arterial blood pressure variations contribute very little to the cuff pressure trajectory. As a matter of fact, the height of those pulses above the cuff pressure baseline is at their maximum when the baseline pressure is equal to the mean arterial pressure (MAP). We demonstrate this in Fig. 3, with a plot of pulses relative to their baseline pressure (pulse-wave amplitude), against their respective baseline cuff pressures. Please note that only

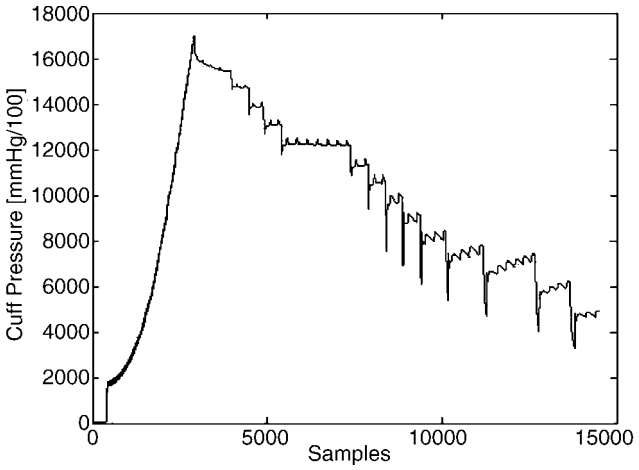


Figure 2. Cuff pressure trajectory when data are collected during step-wise deflation of the cuff (5).

a few of the pulses observed in Fig. 2 are transferred to Fig. 3 to maintain clarity.

Figure 4 shows a cycle of data collected during the continuous inflation of the cuff as well as the pulse-wave amplitude. The pulse-wave amplitude is obtained by subtracting the baseline cuff pressure from the raw pressure data. Next, we will return to the example developed in Figs. 2 and 3 and continue with the estimation of blood pressure values.

It seems trivial to pick the pulse with the tallest height above baseline and to select its baseline pressure to be the MAP. So, for the example at hand, MAP would be just under 100 mmHg (13.33 kPa), as shown in Fig. 5. The systolic and diastolic pressures are then estimated from the MAP using a variety of heuristic rules. A common class of these heuristic rules works as follows. First, the peak values (heights) of the pulse-wave amplitudes are connected to form an envelope. Again, the baseline pressure at the peak of this envelope is the MAP value. Then, the height of the MAP pulse is reduced by a predetermined systolic ratio, and the intersection of this "systolic height" with the envelope to the right of the MAP pulse is selected as the systolic location. The baseline pressure at this location is assigned as the estimate of the systolic pressure, as depicted in Fig. 5. The diastolic pressure is estimated in a similar fashion by using a ratio of its own to arrive at the

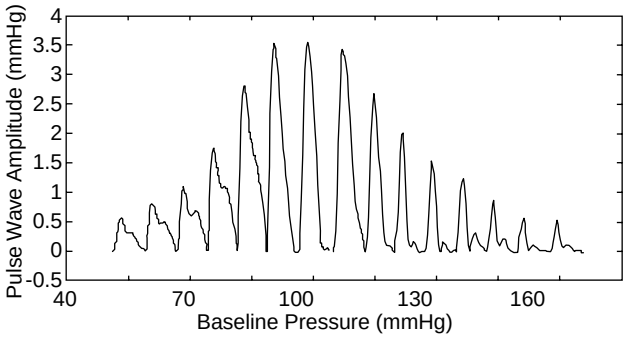


Figure 3. Pulse-wave amplitude profile (6).

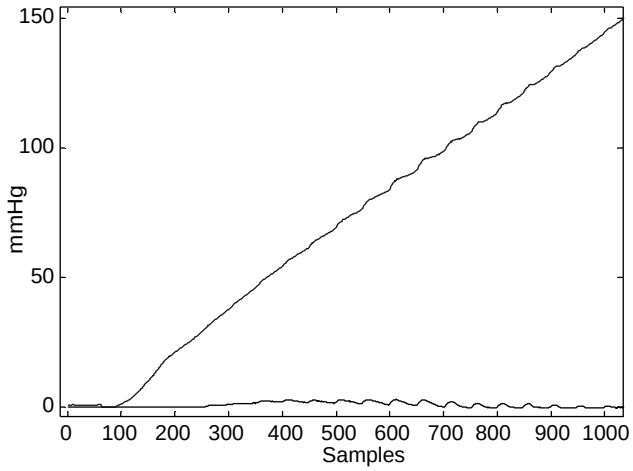


Figure 4. Cuff pressure trajectory and pulse-wave amplitude when data are collected during continuous inflation of the cuff.

diastolic height and then by finding the corresponding intersection with the envelope to the left of the MAP pulse.

In the example shown in Fig. 5, the systolic ratio and diastolic ratio were arbitrarily selected as 0.5 and 0.7, respectively. In a realistic system, those ratios would be found statistically (using methods such as regression, fuzzy rule-based systems, neural networks, or evolutionary algorithms) to minimize deviations between estimated and actual blood pressure values.

Algorithmic Components of Blood Pressure Measurement

In the earlier measurement units, it was a combination of hardware and software that controlled the various aspects of the automated measurement (or estimation) of blood pressure. With the ever increasing computational power of microcontrollers, all decision making and control are now implemented in software and with more elaborate algorithms. Here are some functions that are included in a measurement system. Please refer to Fig. 6 for a typical organization of such algorithms in a blood pressure measurement system.

- **Inflation/deflation control**: Whether data collection is done during inflation or deflation, continuously

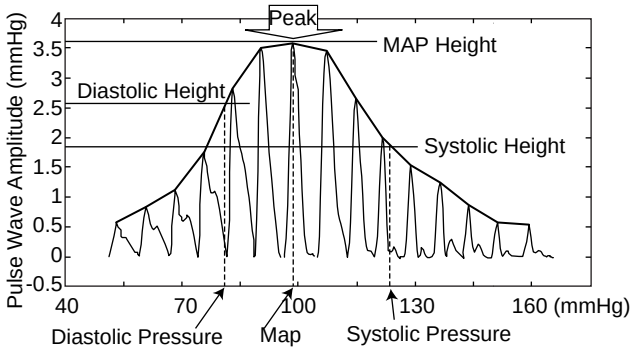


Figure 5. Blood pressure estimation from pulse-wave amplitude profile.

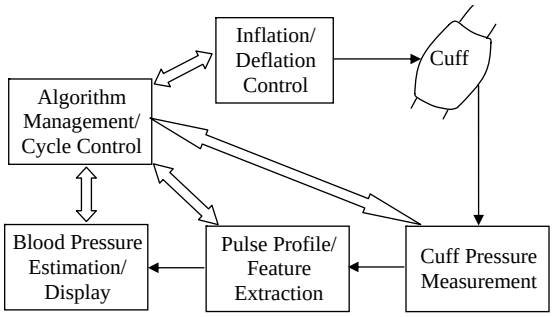


Figure 6. A typical organization of algorithmic components of (oscillatory) blood pressure measurement.

or in steps, there are many challenges to appropriately controlling the air pump. They include maintaining a smooth baseline cuff pressure without filtering out the arterial variations; adjusting the pump speed to variations arising from different cuff sizes, arm sizes, and cuff tightness; and selecting the range of cuff pressures for which data will be collected.

- **Pulse detection**: This is a fundamental part of extracting features from raw cuff pressure data. It becomes especially challenging when conditions such as arrhythmia, or tremors, affect the regularity of pulses. Pattern recognition techniques with features found in time, frequency, or wavelet domains are used to deal with difficult situations.
- **Blood pressure estimation**: The indirect method of measurement is a process of estimating pressures with the use of features extracted from cuff-pressures or other transducer data. This algorithm used to be limited to linear interpolation, as described in the example of Fig. 5. Recently, more elaborate decision-making and modeling tools such as nonlinear regression, neural networks, and fuzzy logic also are being used for this purpose.

Sources of Inaccuracy

Many factors contribute to the inaccuracies in the automated measurement of blood pressure. The following are some of the more significant sources of error:

- **Sparseness of data**: An important design criterion of a blood pressure monitor is to go through a cycle as quickly as possible. However, the faster a device functions, the fewer pulses it will have in a cycle. A cycle time of 1 min would yield about 60–70 pulses, whereas a 20-min cycle would have only 20–23 pulses. The oscillatory techniques are based on collecting cuff pressure due to pulses at baseline pressures that change from above systolic to below diastolic. If we divide a cuff pressure range of about 150–180 mmHg (10.99–23.99 kPa) by the number of pulses in a cycle, we can see that the baseline increment between successive pulses varies from 2 to 3 mmHg (0.26 to 0.39 kPa) in a 1 min cycle to 6 to 9 mmHg (0.79 to 1.19 kPa) in a 20 s cycle. This quantization error affects the accuracy in the estimate of the mean arterial pressure as well as the shape of the pulse envelope, hence, the accuracy of the systolic and diastolic values. Various curve-fitting and interpolation techniques are used to remedy this problem.
- **Pulse extraction uncertainty**: Whether the baseline cuff pressure is varied continuously or in steps, figuring out where one pulse ends and another one starts is not a trivial matter. An inspection of Fig. 2 will show that many artifacts in the data stream may confuse a pulse extraction algorithm and cause errors in the pulse-wave amplitude profile in Fig. 3. In addition, common factors such as an irregularity in the pulses as in arrhythmia, small wrinkles, or folds in the cuff changing its volume suddenly during data collection, or small movements of the patient may amplify those artifacts. A variety of pattern recognition techniques are employed to improve the accuracy of pulse detection (7).
- **Motion artifacts**: The performance of the oscillatory techniques depends on all measurements during a cycle. Therefore, any error caused by a motion of the patient may affect the accuracy of the blood pressure estimations. A comparative study of six noninvasive devices has found that average percent errors due to motion artifacts may be as high as 39% (8). Remedies to this source of error may be a combination of three strategies: (*1*) to identify and compensate for minor artifacts, (*2*) to identify and discard data that include significant artifacts or to repeat the entire cycle if the estimates are deemed unreliable, and (*3*) to incorporate features from additional sensors or monitors such as electrocardiogram (EKG) to help identify motion artifacts (8,9).

Other Blood Pressure Measurement Techniques

Oscillometry is by far the most common technique in automatic noninvasive blood pressure measurement. However, other methods are found in commercial units or in units that are being developed. In this section, a few of these methods are summarized and references are given for further information. It should be noted that algorithmic components and sources of inaccuracy presented within the context of oscillatory technique may apply to other automated measurement methods.

Arterial Tonometry. This relatively new technique in blood pressure measurement is inspired by the tonometry devices that were made in the mid-1950s to measure intraocular pressure. The arterial tonometry device is based on a pressure sensor and pneumatic actuator combination, which is placed on the wrist, above the radial artery. When the pressure applied on the artery is adjusted to the appropriate level (called the hold-down pressure), the portion of the artery wall that is facing the actuator is partially flattened. This configuration maximizes the energy transfer between the artery and the sensor, yielding pulses with the highest amplitude. The relative amplitudes of the tonometry pulses are calibrated to the systolic and the diastolic pressures. Tonometry is suitable for continuous monitoring applications. Sensor placement sensiti-

vity, calibration difficulties, and motion sensitivity are problems that need improvement (10,11).

Pulse-Wave Velocity. A pulse wave is generated by the heart as it pumps blood, and it travels ahead of the pumped blood. By solving analytical equations of fluid dynamics, it has been shown that changes in blood pressure heavily depend on changes in pulse-wave velocity. Blood pressure can be continuously calculated from pulse wave velocity, which in turn is calculated from EKG parameters and peripheral pulse wave measured by an SpO_2 probe on the finger or toe. This method is suitable for continuous monitoring as well as for detecting sudden changes in blood pressure to trigger an oscillometric cycle (12).

Plethysmographic Methods. In this method, changes in the blood volume during a cardiac cycle are sensed using a light emitter and receiver at the finger. Tissue and blood have different infrared light absorbance characteristics. That is, the tissue is practically transparent to the infrared light, whereas blood is opaque to it. A prototype of a ring-like sensor/signal processor/transmitter combination has been reported (13,14)

DIFFERENT FORMS OF BLOOD PRESSURE MEASUREMENT DEVICES

The techniques, algorithms, and transducers discussed in the previous sections have led to a variety of forms of devices, differentiated by where in the body the measurements are taken, or for what purpose the device is used.

Ambulatory Blood Pressure Monitoring

These portable and wearable devices monitor the patient's blood pressure over a long period, say for 24 h. While the patient is following her daily routine, the device periodically takes measurements and saves the results. These measurements are later downloaded for analysis by a physician. The first ambulatory devices, introduced in the early 1960s, were rudimentary and used tape recorders to capture the Korotkoff sounds with an occluding cuff. Most current ambulatory devices use the oscillatory technique. As the patient is subjected to repeated blood pressure measurements with an ambulatory device, it is essential to improve motion tolerance, patient comfort, measurement time, and of course overall accuracy of measurement algorithms that are employed in ambulatory monitors (15).

Ambulatory devices have been instrumental in clinical research and practice. Through their use, there have been significant improvements in our understanding of blood pressure dynamics in a variety of physiological and psychological conditions, and concepts such as "white-coat hypertension," "episodic hypertension," and "circadian rhythm of blood pressure" (e.g., daytime/nighttime variations of blood pressure) have been investigated and added to the medical lexicon (16).

Wrist Blood Pressure Monitoring

These monitors have smaller cuffs than their upperarm-attached counterparts. Hence, they are more compact and more conducive to self-measurement. It is important that the monitors are held at the heart level for correct measurement. They are popular with the home users but typically less accurate than the full-size arm monitors.

Finger Blood Pressure Monitoring

Finger monitors are not nearly as common as the arm or wrist monitors. The approaches used are auscultatory and plethysmographic.

Semiautomatic Blood Pressure Monitoring

The semiautomatic devices have cuffs that are inflated manually by an attached bulb, like a sphygmomanometer. Once the cuff is inflated, the monitor functions in the same manner as an automatic device, taking cuff-pressure measurements while releasing the pressure in a controlled way. These devices are more economical and have longer battery lives than their fully automated counterparts.

ACCURACY OF BLOOD PRESSURE MEASUREMENT DEVICES

Blood pressure measurement devices play an important role in medicine, as they measure one fundamental vital sign. In addition to this traditional use, noninvasive blood pressure devices, especially the automated ones, have become ubiquitous in the home, regularly used by lay people. Two widely used protocols for testing the accuracy of these devices are those set by the Association for the Advancement of Medical Instrumentation (AAMI), a pass/fail system published in 1987 and revised in 1993, and the protocols of the British Hypertension Society (BHS), an A–D graded system, established in 1990 and revised in 1993. These protocols describe in detail the process manufacturers should follow in validating the accuracy of their devices. Their numeric accuracy thresholds can be summarized as follows. A device would pass the AAMI protocols if its measurement error has a mean of no >5 mmHg (0.66 kPa) and a standard deviation of no >8 mmHg (1.06 kPa). The BHS protocol would grant a grade of A to a device if in its measurements 60% of the errors are within 5 mmHg, 85% of the errors are within 10 mmHg (1.33 kPa), and 95% within 15 mmHg (1.99 kPa). BHS has progressively less stringent criteria for the grades of B and C, and it assigns a grade D if a device performs worse than C.

The European Society of Hypertension introduced in 2002 the International Protocol for validation of blood pressure measuring devices in adults (17). The working group that developed this protocol had the benefit of analyzing many studies performed according to the AAMI and BHS standards. One of their motivations was to make the validation process simpler, without compromising its ability to assess the quality of a device. They achieved it by simplifying the rules for selecting subjects for the study. Another change was to devise a multistage process that recognized devices with poor accuracy early on. This is a pass/fail process, using performance requirements with multiple error bands.

Whether blood pressure measurement devices are used by professionals or lay people, their accuracy is important.

Table 2. Summary of Accuracy of Blood Pressure Measurement Devices

Device Type	Number Surveyed	Recommended?		
		Yes	Questionable	No
Manual, clinical	4	1	1	2
Auto, clinical	6	3	2	1
Auto, home, arm	20	4	4	12
Auto, home, wrist	4	0	2	2
Ambulatory	50	26	5	19
Total	84	34	14	36

Yet, most devices in the market have not been evaluated for accuracy independently, using the established protocols (18). In their study, O'Brien et al. surveyed published independent evaluations of manual sphygmomanometers, automated devices for clinical use, and automated devices for personal use. If a device was found acceptable by AAMI standards, and received a grade of A or B by BHS standards, for both systolic and diastolic measurements, then it was "recommended". Otherwise it was not recommended. Few studies they surveyed had issues such as specificity, so devices reported in those studies were "questionably recommended."

Table 2 summarizes the result of their survey. It is interesting to note that of the four clinical grade sphygmomanometers, a kind that is highly regarded by health-care providers, only one was "recommended". Overall, the number of devices "not recommended" is more than the number of "recommended" devices. What one should take away from this analysis is that at every level of quality, price, and target market, it is essential to research the accuracy of a device before investing in it and relying on it.

BIBLIOGRAPHY

Cited References

1. Gibbs NC, Gardner RM. Dynamics of invasive pressure monitoring systems: Clinical and laboratory evaluation. Heart Lung 1988;17:43–51.
2. Webster JG, editor. Medical Instrumentation: Application and Design, 3rd ed. New York: Wiley; 1998.
3. Hambly P. Measuring the blood pressure. Update Anaesthesia 2000;11(6).
4. Philips Invasive Monitoring literature. Available at http://www.medical.philips.com/main/products/patientmonitoring/products/invasivepressure/.
5. Colak S, Isik C. Blood pressure estimation using neural networks. IEEE International Conference on Computational Intelligence for Measurement Systems and Applications, Boston, July 2004.
6. Colak S, Isik C. Fuzzy pulse qualifier. 23rd International Conference of the North American Fuzzy Information Processing Society (NAFIPS 2004) Proceedings, Banff, June 2004.
7. Osowski S, Linh TH. ECG beat recognition using fuzzy hybrid neural network. IEEE Trans Biomed Eng 2001;48:1265–1271.
8. Revision Labs, Beaverton, OR, Noninvasive Blood Pressure Measurement and Motion Artifact: A Comparative Study, December 3, 1998. Available at http://www.monitoring.welchallyn.com/pdfs/smartcufwhitepaper.pdf.
9. Dowling Jr NB. Measuring blood pressure in noisy environments. US patent No. 6,258,037 B1, July 10, 2001.
10. Sato T, Nishinaga M, Kawamoto A, Ozawa T, Takatsuji H. Accuracy of a continuous blood pressure monitor based on arterial tonometry. Hypertension 1993;21:866–874.
11. Matthys K, Verdonck P. Development and modelling of arterial applanation tonometry: A review. Technol Health Care 2002;10:65–76.
12. Williams B. Pulse wave analysis and hypertension: Evangelism versus skepticism. J Hypertension 2004;22:447–449.
13. Yang BH, Asada HH, Zhang Y. Cuff-less continuous monitoring of blood pressure, d'Arbeloff Laboratory of Information Systems and Technology, MIT, Progress Report No. 2–5, March 31, 2000. Available at http://darbelofflab.mit.edu/ProgressReports/HomeAutomation/Report2-5/Chapter01.pdf.
14. Rhee S, Yang BH, Asada HH. Artifact-resistant power-efficient design of finger-ring plethysmographic sensors. IEEE Trans Biomed Eng 2001;48:795–805.
15. McGrath BP. Ambulatory blood pressure monitoring. Med J Australia 2002;176:588–592.
16. National High Blood Pressure Education Program (NHBPEP) Working Group Report On Ambulatory Blood Pressure Monitoring. NIH Publication 92-3028. Reprinted February 1992. Available at http://www.nhlbi.nih.gov/health/prof/heart/hbp/abpm.txt.
17. O'Brien E, Pickering T, Asmar R, Myers M, Parati G, Staessen J, Mengden T, Imai Y, Waeber B, Palatini P. Working Group on Blood Pressure Monitoring of the European Society of Hypertension International Protocol for validation of blood pressure measuring devices in adults. Blood Pressure Monitoring 2002;7:3–17. Available at http://www.eshonline.org/documents/InternationalPS2002.04.29.pdf.
18. O'Brien E, Waeber B, Parati G, Staessen J, Myers MG. Blood pressure measuring devices: Recommendations of the European Society of Hypertension. Br Med J 2001;398. Available at http://bmj.bmjjournals.com/cgi/content/full/322/7285/531.

Further Reading

O'Brien E, Atkins N, Staessen J. State of the market: A review of ambulatory blood pressure monitoring devices. *Hypertension* 1995;26:835–842.

U.S. Food And Drug Administration. Non-Invasive Blood Pressure (NIBP) Monitor Guidance. March 10, 1997. Available at http://www.fda.gov/cdrh/ode/noninvas.html.

See also ARTERIES, ELASTIC PROPERTIES OF; BLOOD PRESSURE, AUTOMATIC CONTROL OF; CAPACITIVE MICROSENSORS FOR BIOMEDICAL APPLICATIONS; LINEAR VARIABLE DIFFERENTIAL TRANSFORMERS.

BLOOD PRESSURE, AUTOMATIC CONTROL OF

YIH-CHOUNG YU
Lafayette College
Easton, Pennsylvania

INTRODUCTION

Arterial pressure is one of the vital indexes of organ perfusion in human bodies. Generally speaking, blood pressure is determined by the amount of blood the heart pumps and the diameter of the arteries receiving blood from the heart. Several factors influence blood pressure. The

nervous system helps to maintain blood pressure by adjusting the size of the blood vessels, and by influencing the heart's pumping action. The heart pumps blood to make sure a sufficient amount of blood circulates to all the body tissues for organ perfusion. The more blood the heart pumps and the smaller the arteries, the higher the blood pressure is. The kidneys also play a major role in the regulation of blood pressure. Kidneys secrete the hormone rennin, which causes arteries to contract, thereby raising blood pressure. The kidneys also control the fluid volume of blood, either by retaining salt or excreting salt into urine. When kidneys retain salt in the bloodstream, the salt attracts water, increasing the fluid volume of blood. As a higher volume of blood passes through arteries, it increases blood pressure.

Hypertension is defined as abnormal high systemic arterial blood pressure, systolic and diastolic arterial pressures > 140 and 95 mmHg (18.662 and 12.664 kPa). The causes of hypertension might be due to acute myocardial infarction, congestive heart failure, and malignant hypertension. Postoperative cardiac patients may experience hypertension because of pain, hypothermia, reflex vasoconstriction from cardiopulmonary bypass, derangement of the rennin-angiotension system, and ventilation difficulties. A prolonged postoperative hypertension could lead to complications, including myocardial ischemia, myocardial infarction, suture line rupture, excessive bleeding, and arrhythmia. As a result, clinical treatment to postoperative hypertension is needed to reduce the potential risk of complications.

Postoperative hypertension is usually treated pharmacologically in the intensive care unit (ICU). Sodium nitroprusside (SNP) is one of the most frequently used pharmaceutical agents to treat hypertensive patients and is a vasodilating drug that can reduce the peripheral resistance of the blood vessel, and thus causes the reduction of arterial blood pressure. A desired mean arterial pressure (MAP) can be achieved by monitoring MAP and regulating the rate of SNP infusion. The mean arterial pressure can be measured from a patient by using an arterial pressure transducer with appropriate signal amplification. Low pass filtering is used to remove high frequency noise in the pressure signal and provide MAP for monitoring purpose. Administration of SNP infusion could be performed by manual operation. The drug infusion rate should be adjusted frequently in response to the spontaneous pressure variation and patient's condition changes. In addition, blood pressure response to the drug infusion changes over time and varies from patient to patient. Therefore, this manual approach is extremely difficult and time consuming for the ICU personnel. As the result, the use of control techniques to regulate the infusion of the pharmaceutical agents and maintain MAP within a desired level automatically has been developed in the last 30 years.

IVAC Corporation developed an automatic device, TITRATOR, to infuse SNP and regulate MAP in postoperative cardiac patients in early 1990s. Clinical evaluation for the clinical impact of this device in multiple centers was reported by Chitwood et al. (1). Patients who participate in this trial were treated by either automatic or manual control. The automated group showed a significant reduction in the number of hypertensive episodes per patient. Chest tube drainage, percentage of patients receiving transfusion, and total amount transfused were all reduced significantly by the use of an automated titration system. Although TITRATOR was not commercialized successfully due to economic reasons, the promising clinical experiences encouraged future development of automatic blood pressure regulation devices.

An automatic blood pressure control system usually includes three components: sensors, a controller, and a drug delivery pump. This article provides an overview of automatic control schemes, including proportional-integral-derivative (PID) controllers, adaptive-controllers, rule-based controllers, and artificial neural network controllers that regulate mean arterial blood pressure using SNP. A brief description of each control strategy is provided, followed by examples from literature. Testing of the control performance in computer simulations, animal studies, and clinical trials, is also discussed.

CONTROL SCHEMES

PID Controller

The PID control of MAP determines the SNP infusion rate, $u(t)$, based on the difference between the desired output and the actual output,

$$u(t) = K_{\mathrm{P}} e(t) + K_{\mathrm{I}} \int_{t_0}^{t_1} e(t)dt + K_{\mathrm{D}} \frac{d}{dt} e(t) \qquad (1)$$

where $e(t) = P_{\mathrm{d}}(t) - P_{\mathrm{m}}(t)$, $P_{\mathrm{d}}(t)$ is the desired MAP, and $P_{\mathrm{m}}(t)$ is the actual mean arterial pressure. The parameters K_{P}, K_{I}, and K_{D} are the proportional, integral, and differentiation gain respectively. The design of this type of controller involves the selection of appropriate control gains, K_{P}, K_{I}, and K_{D}, such that the actual blood pressure, $P_{\mathrm{m}}(t)$, can be stabilized and maintained close to the desired level, $P_{\mathrm{d}}(t)$. Typical components of the automatic blood pressure control system, including the PID controller, the infusion pump, the patient, as well as the patient monitor along with physiologic sensors are illustrated in Fig. 1.

Sheppard and co-worker (2–4) developed a PI-type controller, by setting $K_{\mathrm{D}} = 0$ in (1), to regulate SNP, which has been tested over thousands of postcardiac-surgery patients in the ICU. The control gains were tuned to satisfy an acceptable settling time with minimal overshoot. The discrete-time PI controller updates the infusion rate as

$$u(k) = u(k-1) + \Delta u(k) \qquad (2)$$

where $u(k-1)$ is the previous infusion rate a minute ago and $\Delta u(k)$ is the infusion rate increment defined by,

$$\Delta u(k) = K\{0.4512\, e(k) + 0.4512\, [e(k) - e(k-1)]\} \qquad (3)$$

where $e(k)$ and $e(k-1)$ are the current and previous error, respectively. The gain K in Eq. 3 as well as the further correction of $\Delta u(k)$ were determined by the region of current MAP $P_m(k)$ as described in the following:

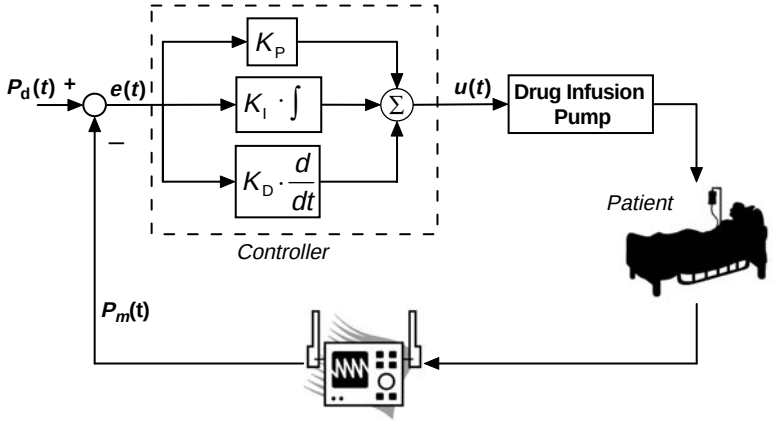


Figure 1. Proportional-integral-derivative control scheme of blood pressure control.

Rule 1 If $P_m(k) \geq P_d + 5$, then $K = -1$ and $\Delta u(k) = \Delta u(k)$ from Eqs. 3–2

Rule 2 If $P_d \leq P_m(k) < P_d + 5$, then $K = -0.5$ and $\Delta u(k) = \Delta u(k)$ from Eq. 3

Rule 3 If $P_d - 5 \leq P_m(k) < P_d$, then $K = -1$ and $\Delta u(k) = \Delta u(k)$ from Eq. 3

Rule 4 If $P_m(k) \leq P_d - 5$, then $K = -2$ and $\Delta u(k) = \Delta u(k)$ from Eq. 3

Rule 5 If $P_m(k) < P_d - 5$ and $\Delta u(k) > 0$, then $\Delta u(k) = 0$

Rule 6 If $P_m(k) \geq P_d$ and $\Delta u(k) > 7$, then $\Delta u(k) = 7$

These rules were designed to provide a boundary for the controller and achieve the optimal performance with the minimal pharmacological intervention. As a result, the controller is a nonlinear PI-type controller.

The automatic blood pressure controller described herein performed better than human operation in a comparison study (5). Automatic blood pressure regulation exhibits approximately one-half of the variation observed during manual control; MAP are more tightly distributed about the set-point, as shown in Fig. 2 (2). Forty-nine postcardiac surgery patients in ICU were managed by the automatic controller. The patients' MAPs were maintained within ± 5 mmHg (± 0.667 kPa) of the desired MAP 94% of the total operation time (103 out of the 110 operation hours). A group of 37 patients were managed with manual operation provided by experienced personals, with which only 52% of the time the patients' MAPs were within the prescribed range.

Adaptive Controller

The PID controller considered previously was with the control gains determined prior to their implementation. The control gains were usually tuned to satisfy the performance criterion in simulation or animal studies where the parameters characterizing the system dynamics were fixed variables. In clinical applications, the cardiovascular vascular dynamics change over time as well as from patient to patient. In addition, the sensitivity to drugs varies from one patient to another and even with the same patient at different instant. Therefore, it would be beneficial if the control gains can be adjusted automatically during operation to adapt the differences between patients as well as physiologic condition changes in a patient over time. This type of controllers is called adaptive controller.

An adaptive control system usually requires a model, representing plant (the patient and the drug infusion system) dynamics. Linear black box models, expressed by

$$\begin{aligned} y(k) &= \frac{B(q^{-1})}{A(q^{-1})}u(k) + \frac{C(q^{-1})}{A(q^{-1})}n(k) \\ A(q^{-1}) &= 1 + a_1 q^{-1} + a_2 q^{-2} + \cdots + a_n q^{-n} \\ B(q^{-1}) &= 1 + b_1 q^{-1} + b_2 q^{-2} + \cdots + b_l q^{-l} \\ C(q^{-1}) &= 1 + c_1 q^{-1} + c_2 q^{-2} + \cdots + c_m q^{-m} \end{aligned} \tag{4}$$

are typically used to represent the plant dynamics. A, B, and C are polynomials in the discrete shift operator q, where a_i, b_i, and c_i are coefficients in the polynomials; $y(k)$, $u(k)$, and $n(k)$ are the model input, output, and noise, respectively. Depending on the polynomials B and C, the model in Eq. 4 can be classified as autoregressive [AR, $B(q^{-1}) = 0$, $C(q^{-1}) = 1$], autoregressive with inputs [ARX, $C(q^{-1}) = 1$], autoregressive moving average [ARMA, $B(q^{-1}) = 0$], and autoregressive moving average with inputs (ARMAX). The coefficients of the polynomials are time-varying, much slower than the plant dynamic changes. The controller updates the control input, $u(k)$, by taking the model parameter changes into consideration. General reviews and descriptions on adaptive control theory can be found in literature (6–8). Three types of adaptive control schemes are frequently used in blood pressure controller design: self-tuning regulator, model reference adaptive control, and multiple model adaptive control.

Self-Tuning Regulator. The self-tuning regulator (STR) is based on the idea of separating the estimation of unknown parameters from the design of the controller. It is assumed that a priori knowledge of the model structure, that is, l, m, and n in Eq. 4. In choosing l, m, and n, one must compromise between obtaining an accurate representation of the system dynamics while keeping the system representation simple. The parameters of the regulator are adjusted by using a recursive parameter estimator and a

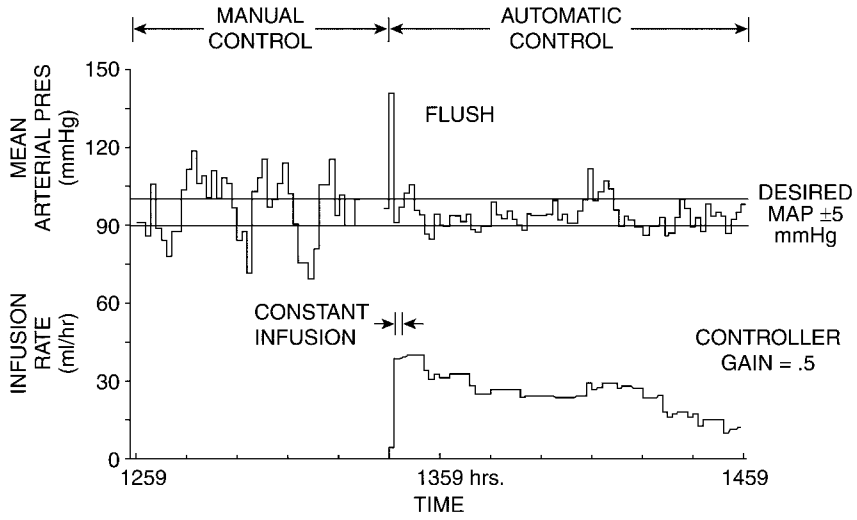


Figure 2. Comparison of manually controlled SNP infusion with computer control in the same patient. (Redrawn with permission from L.C. Sheppard, Computer control of the infusion of vasoactive drugs, *Ann. of Biomed. Eng.*, Vol. 8: 431–444, 1980. Pergamon Press, Ltd.)

regulator design calculation as shown in Fig. 3. The parameter estimates are treated as if they are true or at least asymptotically the true parameters. Several algorithms are available for parameter estimation, including recursive least-squares, generalized least-squares, stochastic approximation, maximum likelihood, instrumental variables, and Kalman filter. Every technique has its advantages and disadvantages. Descriptions of parameter estimation algorithms can be found in (9). Various approaches are available for regulator design calculation, such as minimum variance, gain and phase margin analysis, pole placement, and linear quadratic Gaussian (LQG). More detailed information of STR can be found in literature (6–8).

Various STR-type blood pressure controllers have been developed and tested in computer simulations, animal experiments, as well as clinical studies. Arnsparger et al. (10) used a second-order ARMA model to design the STR. A recursive least-mean-squares estimator was used to estimate the model parameters. The parameter estimates were then used to calculate the control signal, the drug infusion rate, based upon a minimum variance or a one-step-ahead control law. Both algorithms were implemented in microprocessor and tested in dog experiments for comparison. Both controllers were able to maintain the MAP at the desired level. However, the one-step-ahead controller performed better in the test with less variation in the infusion rate.

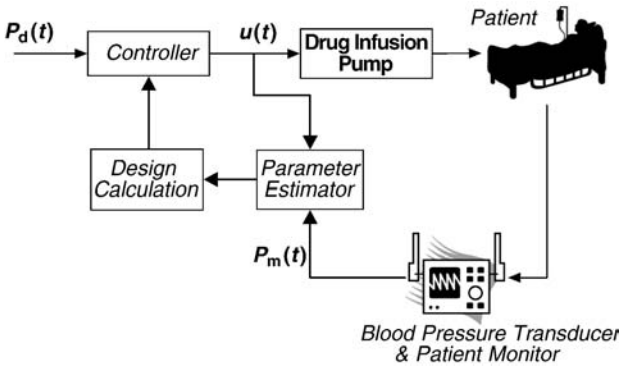


Figure 3. Configuration of self-tuning regulator for blood pressure control.

A combination of proportional derivative with minimum variance adaptive controller was designed by Meline et al. (11) to regulate MAP using SNP. The plant dynamics was represented by a fifth-order ARMAX model, while the model parameters were estimated through a recursive least-squares algorithm. The controller was tested on ten dog experiments as well as human subjects (12). Twenty patients with postsurgical hypertension were randomly assigned to either the manual group, where SNP was administrated by experience nurse, or the automatic group. Statistical analysis showed that MAP was maintained within $\pm$ 10% from the desired MAP for 83.3% of the total operation time in the "automatic" group versus 66.1% of the total operation time in the "manual" group. This implies the automatic control performed better than the manual operation.

A pole-assignment STR was designed by Mansour and Linkens (13) to regulate blood pressure using a fifth-order ARMAX model. The model parameters were identified through a recursive weighted least-squares estimator. These parameters were then used to determine appropriate feedback gains for the controller. Pole-placement algorithm was used because of its robustness to a system with nonminimum phase behavior or unknown time delay. Effectiveness of the controller was evaluated extensively in computer simulation, using a clinically validated model developed by Slate (3) as shown in Fig. 4(2). The controller demonstrated a robust performance even with the inclusion of the recirculation term or a variable time delay.

Voss et al. (14) developed a control advance moving average controller (CAMAC) to simultaneously regulate arterial pressure and cardiac output (CO) using SNP and dobutamine. CAMAC is a multivariable STR, which has the advantage of controlling nonminimum phase plants with unknown or varying dead times. The controller determines the drug infusion rates based on the desired MAP and CO, past inputs, past outputs, and a on-line recursive least-squares estimator with an exponential forgetting factor identifying the subject's response to the drugs.

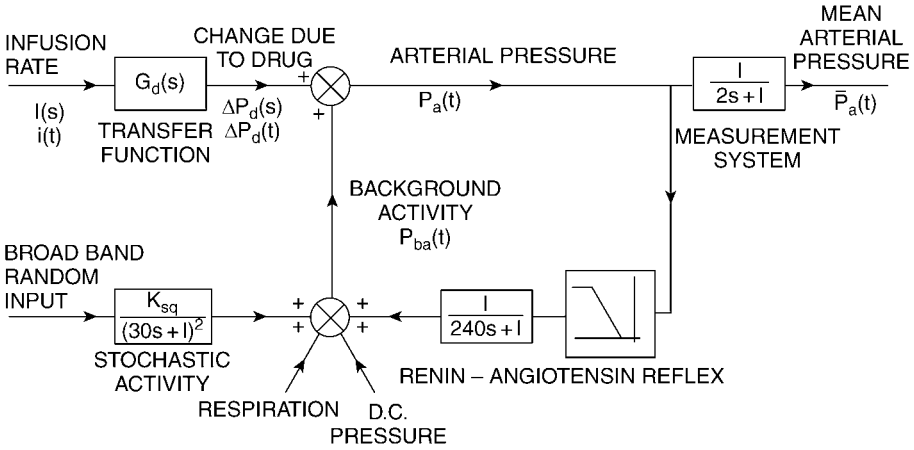


Figure 4. Model of MAP in response to SNP infusion. (Redrawn with permission from L.C. Sheppard, Computer control of the infusion of vasoactive drugs, *Ann. Biomed. Eng.* 1980; 8: 431–444. Pergamon Press, Ltd.)

The plant model for designing the controller and estimator was a second-order ARMAX model. The control algorithm was designed and tested in simulations prior to dog experiments. Although animal studies demonstrated that the controller was capable to maintain MAP and CO at their desired level, changing vasomotor tone and the lack of high frequency excitation signals could lead to inaccuracy in the parameter estimation, causing poor performance in transient response.

Model Reference Adaptive Control. The basic principle of the model reference adaptive control is illustrated in Fig. 5). The desired input–output response is specified by the reference model. The parameters of the regulator are adjusted by the error signal, the difference between the reference model output and the system output, such that the system output follows the reference output. More detailed information about MRAC can be found in Ref. 7.

The use of MRAC to regulate blood pressure was introduced by Kaufmann et al. (15). The format of the reference model was adopted form that developed by Slate (3). Controller design and evaluation were carried out in computer simulation. The controller with adaptation gains showed lower steady-state error than that with nonadaptive gains in simulations, particularly when a process disturbance was introduced. Animal studies were conducted to compare the performance of the MRAC with that of a well-tuned PI controller. Neosynephrine was introduced to change the transfer function characteristics of the subjects during experiments. The MRAC was superior to the PI controller and maintained MAP closed to the reference with an error within ± 5 mmHg (± 0.667 kPa) regardless of the plant characteristic changes due to drug intervention.

Pajunen et al. (16) designed a MRAC to regulate blood pressure using SNP with the ability to adjust the reference model by learning the patient's characteristics, represented by the model parameters, coefficients and time delays, of the transfer function. These model parameters were assumed to be unknown and exponentially time-varying. The time-varying reference model was automatically tuned to achieve the optimal performance while meeting the physical and clinical constraints imposed on the drug infusion rate and MAP. Extensive computer simulation was used to evaluate the robustness of the controller. The MAP was maintained within ± 15 mmHg (± 2 kPa) around the set-point regardless of changes in patient's characteristics and the presence of high level noises.

Polycarpou and Conway (17) designed a MRAC to regulate MAP by adjusting SNP infusion rate. The plant model was a second-order model discretized from the Slate's model (3). Time delay terms in the model were assumed to be known while the model parameters were constant with nonlinear terms. The constant terms were assumed to be known and the nonlinear terms were estimated by a radial basis function (RBF) neural network. The resulting parameter estimates were then used to update the control law such that the system output follows the reference model. Although the RBF was able to model the unknown nonlinearity and thus improve the closed-loop characteristics in computer simulation, the

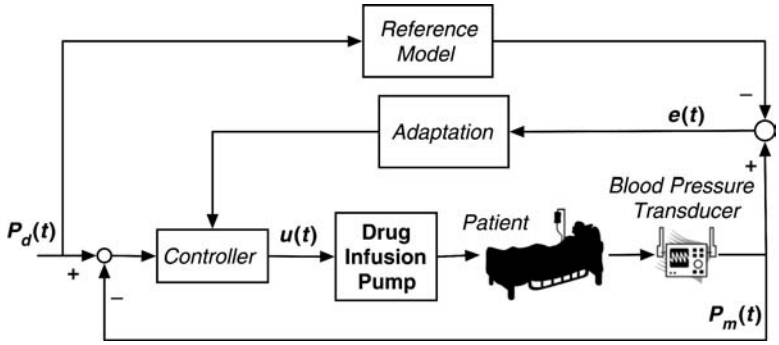


Figure 5. Configuration of MRAC for blood pressure regulation.

assumption that the model parameters and time delays were known would need further justification in practical applications.

Multiple Model Adaptive Control. The concept of multiple model adaptive control (MMAC) was first introduced by Lainiotis (18). This technique assumes that the plant response to the input can be represented by a bank of models. A controller is designed *a priori* to give a specified performance for each particular model. A probability, $P(q_i|t)$, describing the accuracy of each model, q_i, to represent the actual system, is calculated and used as the weighting factor to update the control input,

$$u = \sum_{i=1}^{N} u_i \cdot P(q_i|t) \tag{5}$$

where u_i is the control input based on the model q_i. As the response of the system changes, the probability, $P(q_i|t)$, will also be adjusted accordingly such that the model closest represents current dynamics gets the greatest probability. As a result, the contribution of the control input, obtained from the model with the greatest probability, to the updated control input in equation 5 is more significant than the inputs from other models with lower probabilities. Configuration of the MMAC is illustrated in Fig. 6.

He et al. (19) introduced the first blood pressure controller using the MMAC technique. There were eight plant models derived from Slate's model (3) for controller design. Each plant model contains a constant model gain between 0.32 and 6.8, representing the plant gain of 0.25–9 in Slate's model (3), along with the same time constants and delays at their nominal values. A proportional-plus-integral (PI) type controller was designed for each plant model. These controllers were with the same time constant but different gains. Computer simulation was used to test the controller performance in response to the variations of model parameters and the presence of background noise. The controller was able to settle MAP within 10 min with the error within ± 10 mmHg (± 1.333 kPa) from the set-point. The control algorithm was further tested in animal experiments. The controller stabilized MAP in < 10 min with ± 5 mmHg (± 0.667 kPa) error from its set-point, regardless of the plant characteristic changes due to neosynephrine injection, the sensitivity of the subject to the SNP infusion, and the background noise. The mean error was < 3 mmHg (0.4 kPa) over the entire studies.

Martin et al. (20) developed a MMAC blood pressure controller with seven models modified from Slate's model (3). The model gains in the seven models were from 0.33 to 9.03 to cover the variation of the plant gain between 0.25 and 10.86. The other model parameters were held constant at their nominal values. A pole-placement compensator was designed for each model. A Smith predictor was used to remove the effects of infusion delay, and thus simplify the control analysis and design. A PI unit was included to achieve zero steady-state error. Two constrains were used to limit the infusion rate when the patient's blood pressure is too low or the resulting SNP infusion rate from the controller is beyond the preset threshold. The controller was able to maintain MAP with the settling time < 10 min, the maximum overshoot < 10 mmHg (1.333 kPa), and the steady-state error within ± 5 mmHg (± 0.667 kPa) around the pressure set-point in computer simulations. The controller was also tested on 5 dogs as well as 19 patients during cardiac surgery with the aid of a supervisor module, which oversees the overall environment and thus improves the safety (21,22).

Yu et al. (23) designed a MMAC to control MAP and CO by adjusting the infusion rates of SNP and dopamine for congested heart failure subjects. There were 36 linear multiinput and multioutput (MIMO) models, represented by first-order transfer functions with time delays, to cover the entire range of possible dynamics. A model predictive controller [MPC, (24)] was designed for each individual model to find a sequence of control signals such that a quadratic cost function can be minimized. In order to save computation time, only the control signals corresponding to the six models with the highest probability weights were used to determine the drug infusion rates. The control

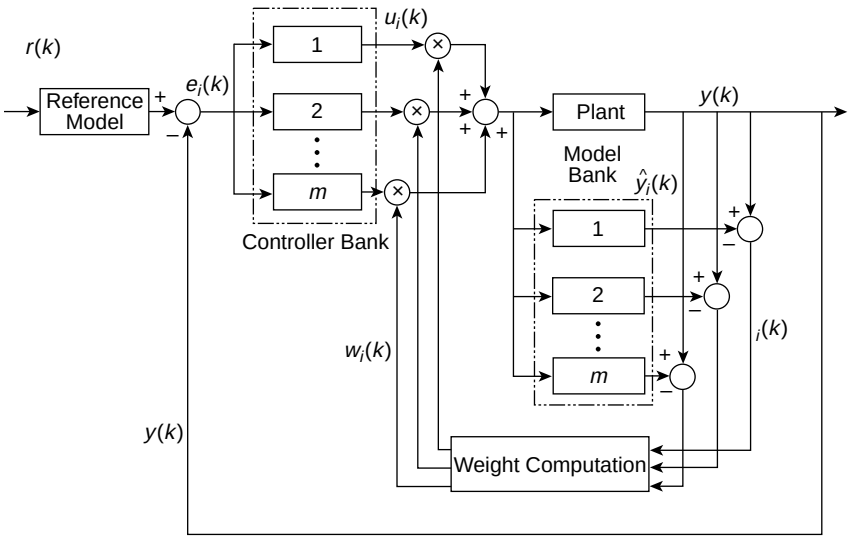


Figure 6. Block diagram of multiple model predictive control. [Redrawn with permission from Rao et al., Automated regulation of hemodynamic variables, *IEEE Eng. Med. Biol.* 2001; 20 (1): 24–38. (© Copyright 2004 IEEE).]

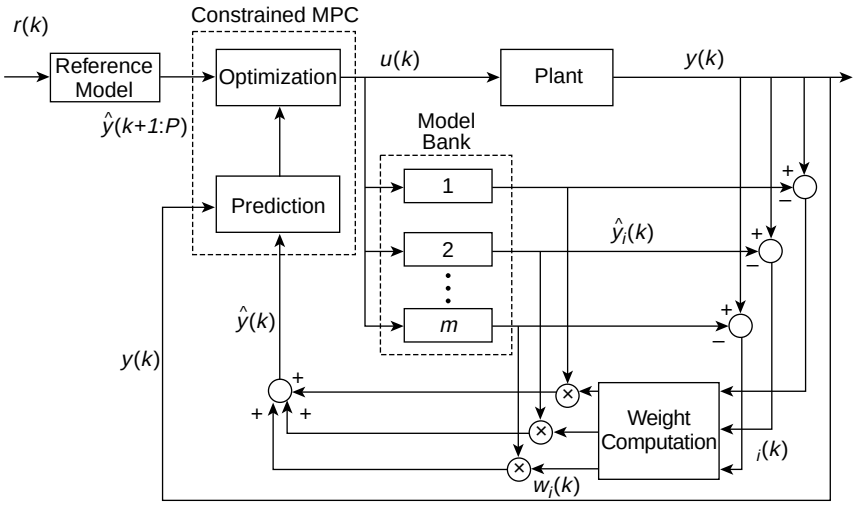


Figure 7. Modified multiple model predictive control strategy. [Redrawn with permission from Rao et al., Automated regulation of hemodynamic variables, *IEEE Eng. Med. Biol*. 2001; 20 (1): 24–38. (© Copyright 2004 IEEE).]

algorithm was tested in six dogs, including some cases with induced heart failure. It took 3–10.5 min to settle MAP within ± 5 mmHg (± 0.667 kPa) of the steady-state set-point with the mean of 5.8 min in all cases. The overshoot was between 0 and 12 mmHg (1.6 kPa) with the average of 5.92 mmHg (0.79 kPa). The standard deviation of MAP about its set-point was 4 mmHg (0.533 kPa).

The major challenge of implementing MPC in MMAC is the computation time, especially for a large model bank. Rao et al. (25) designed a MMAC with a single constrained MPC as shown in Fig. 7. The model bank, constituted first-order-plus-time-delay MIMO models spanning sufficient spectrum of model gains, time constants, and time delays, was run in parallel to obtain the possible input–output characteristics of a patient's response to drug dosages. A Bayesian weight was generated for each model based on the patient's response to drugs. The MPC used the combination of model weights to determine the optimal drug infusion rates. This control scheme combines the advantages of model adaptation according to patient variations, as well as the ability to handle explicit input and output constraint specifications. The controller effectively maintained MAP and cardiac output in seven canine experiments (26). Figure 8 illustrates the results of control MAP and CO using SNP and dopamine in on study. High levels of fluothane were introduced to reduce CO, mimicking congestive heart failure. The controller achieved both set-points of MAP = 60 mmHg (8 kPa) and CO = 2.3 $\text{L}\cdot\text{min}^{-1}$ in $\sim$ 12 min. In average over the entire studies, MAP was maintained within ± 5 mmHg (± 0.667 kPa) of its set-point 89% of the time with a standard deviation of 3.9 mmHg (0.52 kPa). Cardiac output was held within ± 1 $\text{L}\cdot\text{min}^{-1}$ of the set-point 96% of the time with a standard deviation of 0.5 $\text{L}\cdot\text{min}^{-1}$. Manual regulation was performed in the experiments for comparison. The MAP was kept within ± 5 mmHg (± 0.667 kPa) of its set-point 82% of the time with a standard deviation of 5.0 mmHg (0.667 kPa) while CO stayed in the ± 1 $\text{L}\cdot\text{min}^{-1}$ band of the set-point 92% of the time with a standard deviation of 0.6 $\text{L}\cdot\text{min}^{-1}$. Clearly, the automatic control performed better than the manual approach.

Rule-Based Controller

The blood pressure controllers discussed previously rely on mathematical models that can characterize plant dynamics, including the drug infusion system, human cardiovascular dynamics, and pharmacological agents. Identifying such mathematical forms could be a challenge due to the complexity of human body. Despite this, there exist experienced personnel, whose ability to interpret linguistic statements about the process and to reason in a qualitative fashion prompts the question: "can we make comparable use of this information in automatic controllers?"

In rule-based or intelligent control, the control law is generated from linguistic rules. This model-free controller usually consists of an inference engine and a set of rules for reasoning and decision making. A typical control rules are represented by *if* <*condition*> *then* <*action*> statements. Rule-based approaches have been proposed as a way of dealing with the complex natural of drug delivery systems and, more importantly, as a way of incorporating the extensive knowledge of clinical personnel into the automatic controller design.

One of the most popular rule-based control approaches is fuzzy control. Fuzzy control approach is based on fuzzy set theory and is a rule-based control scheme where scaling functions of physical variables are used to cope with uncertainty in the plant dynamics. A typical fuzzy controller, shown in Fig. 9, usually includes three components: (*1*) membership functions to fuzzify the physical input, (*2*) an inference engine with a decision rule base, and (*3*) a defuzzifier that converts fuzzy control decisions into physical control signals. More details on fuzzy set theory and its control applications are available in (27–29).

Isaka et al. (30) applied an optimization algorithm to determine the membership functions of a fuzzy blood pressure controller using SNP. This method reduced the time and efforts to determine appropriate values for a large number of membership functions. In addition, it also provided the knowledge of the effect of membership functions to the fuzzy controller performance, as well as the effect of

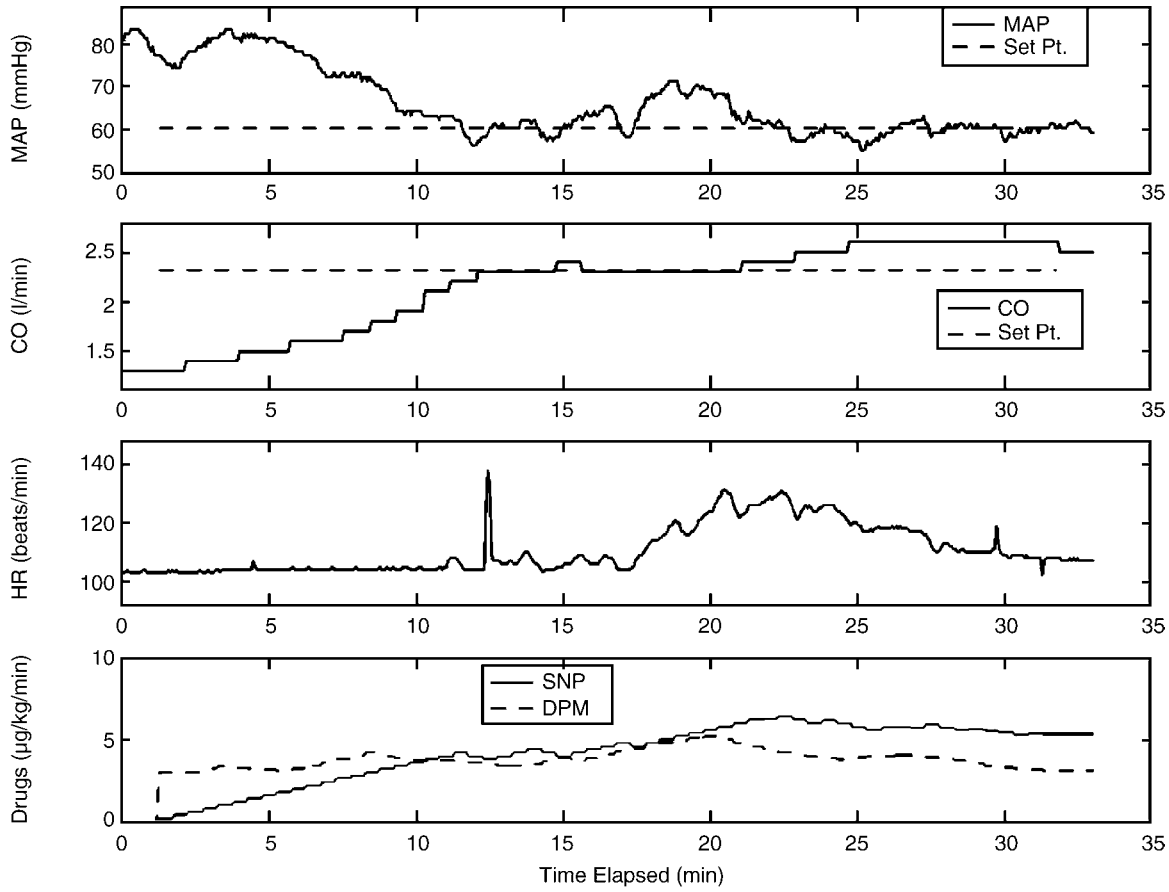


Figure 8. Multiple model adaptive control of MAP and CO using SNP and dopamine in canine experiment. (© Copyright 2004 IEEE).

plant parameter variations to the changes in membership functions. Efficacy of using this controller to regulate MAP by infusing SNP was evaluated in computer simulation model proposed by Slate (3). The MAP was initialized at 120 mmHg (16 kPa) at the beginning of simulation. The target MAP value was first set at 80 mmHg (10.665 kPa) and then changed to 110 mmHg (14.665 kPa). The target MAP values were achieved in < 3 min with overshoots < 10 mmHg (1.333 kPa).

Ying et al. (31) designed an expert-system-shell-based fuzzy controller to regulate MAP using SNP. The controller was a nonlinear PI-type control while the control gains were predetermined by analytically converting the fuzzy control algorithm. This converting process provided the advantage of execution time reduction. The controller was further fine-tuned to be more responsive to the rapid and large changes of MAP. It was successfully tested in 12 postsurgical patients for the total of 95 hs and 13 min. MAP was maintained within $\pm 10\%$ of its target value, 80 mmHg (10.665 kPa), 89.3% of the time over the entire test.

Neural-Network Based Controller

Artificial neural networks (ANN) are computation models that have learning and adaptation capabilities. An ANN-based controller is usually more robust than the traditional controllers in the presence of plant nonlinearity and uncertainty if the controller is trained properly. A survey article about the use of ANN in control by Hunt et al. (32) provides more detailed information.

The use of ANN-type controller in arterial blood pressure regulation was investigated in feasibility studies in either computer simulation or animal experiments. Chen et al. (33) designed an ANN-type adaptive controller to control MAP using SNP. The controller was tested in computer simulation with various gains and different levels of noise. The controller was able to maintain MAP close to the set point, 100 mmHg (13.33 kPa) with error within ± 15 mmHg (± 2 kPa) in an acceptable tolerance settling time < 20 min.

Kashihara et al. (34) compared various controllers, including PID, adaptive predictive control using ANN (APP_{NN}), a combined control of PID with APP_{NN}, a fuzzy controller, and a model predictive controller, to maintain MAP for acute hypotension using norepinephrine. The controllers were tested in computer simulation and animal studies. The controllers based on neural network approach were more robust in the presence of unexpected hypotension and unknown drug sensitivity. Adding an ANN or a fuzzy logic scheme to the PID or adaptive controller improved the ability of the controller to handle unexpected conditions more effectively.

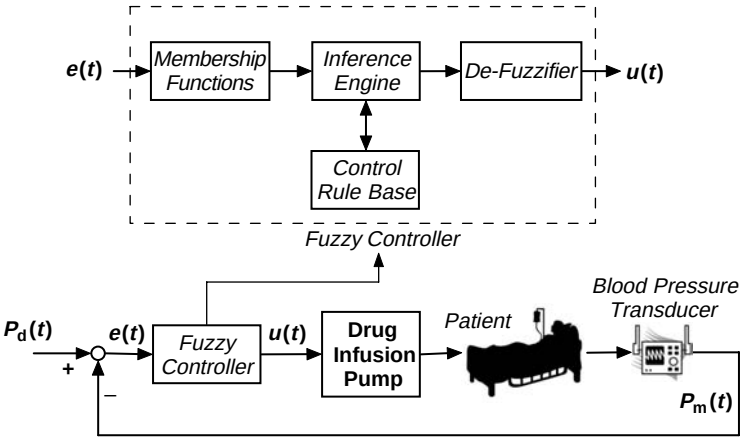


Figure 9. Block diagram of a fuzzy controller. [Redrawn with permission from Isaka et al., An optimization approach for fuzzy controller design, 1992; SMC 22: 1469–1473. (© Copyright 2004 IEEE).]

DISCUSSION

Numerous controllers have been developed since 1970s to regulate SNP and control MAP for hypertension patients. The control strategies can generally be classified as PID control, adaptive control (including STR, MRAC, and MMAC), rule-based control, as well as neural network control. Most controllers were developed and tested in computer simulation and animal experiments successfully. A few controllers were tested clinically with satisfactory results. Table 1 summarizes the control algorithms reviewed in this article.

Controller performance is influenced by several factors, including the fit of the process model to the plant, signal conditioning of the sensors under various clinical environments, as well as the diagnosis ability of the devices. Model selection is crucial for the stability and robustness of a controller. In blood pressure regulation, variable time delay, patient's sensitivity to SNP, and rennin regulatory mechanism are important factors. These factors could cause parameter variations in the plant model that might reduce the performance of a fixed-gain controller. Adaptive controllers that can adjust the control signal based on the estimation of model parameters or the probability of model errors could overcome the limit of the fixed-gain control. Some controllers have been tested in the laboratory with promising results. However, clinical applications of this type of controllers were very few. Rule-based and ANN controllers do not need a specific plant model for control design. The training signals or information must provide a broad coverage of possible events in the clinical environment to assure the reliability of the control algorithm.

Patient care practices and other aspects of the clinical environment must be considered in the design of a clinical useful system. A supervisory algorithm that can detect potential risks, determine appropriate control signals to

Table 1. Summary of Blood Pressure Controllers Reviewed in this Article

		Controller Performance			Controller Test		
Articles	Control Scheme	Settling Time (min)	Overshoot, mmHg	Steady-State about the Set-Point, mmHg	Simulation	Animal Studies	Clinical Studies
Slate et al.(2–4)	Nonlinear PI	<10		± 10	x	x	x
Arnsparger et al.(10)	STR	2	30	10		x	
Mansour et al.(13)	STR	5–20	<10	± 5	x		
Voss et al.(14)	CAMAC	1.3–7.3	0–22	−4–9.8	x	x	
Kaufmann et al.(15)	MRAC (w/known time-constant and delay)	<5		± 5	x	x	
Pajunen et al.(16)	MRAC (w/time-varying parameters)	<5	<15	± 15	x		
Polycarpou et al.(17)	MRAC	5		± 10	x		
He et al.(19)	MMAC	<8	<5	± 5	x	x	
Martin et al.(20–22)	MMAC	<10	<10	± 5	x	x	x
Yu et al.(23)	MMAC	3–10.5	0–12	± 5	x	x	
Rao et al.(25,26)	MMPC	12		± 5	x	x	
Isaka et al.(30)	Fuzzy controller	<3	<10	± 5	x		
Ying et al.(31)	Fuzzy controller			± 8	x		x
Chen et al.(33)	ANN	5 to 20		± 15	x		
Kashihara et al.(34)	ANN	2		± 5	x	x	

stably maintain a patient's blood pressure near the set point, and identify excessive noise or artifact in sensor measurements would be beneficial (21,22,31). The supervisor oversees the entire conditions of the control environment and directs the controller to take control actions efficiently and safely. Control decision is based upon sensor measurements. It is very important that the supervisor is able to process measurements and detect the nonphysiological signals, such as the noisy signals duo to suction the airway and flushing the arterial catheter, and thus avoid acting on unreliable information. In addition, the supervisor must have the ability to assure the proper operation of the infusion system for drug delivery. This monitoring system should be able to detect the potential faults that could prevent abnormal operation of the device (e.g., blood clotting, infusion kinking, leakage, and infusion pump stoppage).

CONCLUSION

Because of the quick action of SNP in blood pressure reduction, frequent monitoring of MAP followed by infusion rate adjustment is necessary. The use of manual control to achieve desired MAP would be burdensome to ICU personnel, who are already loaded with many duties. Successful development of a blood pressure controller that could automatically maintain patient's MAP within a preset range with self-monitoring capability would reduce the workload of the patient care providers and improve the patient's quality of life in the clinical environment.

Blood pressure control systems designed previously provide valuable experiences for further development. The future controller should be able to adapt the characteristic changes (represented by gains, time delays, and time constants) from patient to patient as well as the variations within a patient over time. In order to improve the reliability and safety of the controller, incorporating a supervisory scheme that can monitor system operation as well as identify and manage unexpected mechanical errors and clinical environment changes with the control system would be essential.

BIBLIOGRAPHY

Cited References

1. Chitwood Jr WR, Cosgrove 3rd DM, Lust RM. Multicenter trial of automated nitroprusside infusion for postoperative hypertension. Titrator Multicenter Study Group. Ann Thorac Surg 1992;54:517–522.
2. Sheppard LC. Computer control of the infusion of vasoactive drugs. Ann Biomed Eng 1980;8:431–444.
3. Slate JB. Model-based design of a controller for infusing sodium nitroprusside during postsurgical hypertension. Ph.D. dissertation. University of Wisconsin-Madison, 1980.
4. Slate JB, Sheppard LC. Automatic control of blood pressure by drug infusion. Proc Inst Electr Eng 1982;129, (Pt. A):639–645.
5. de Asla RA, Benis AM, Jurado RA, Litwak RS. Management of postcardiotomy hypertension by microcomputer-controlled administration of sodium nitroprusside. J Thrac Cardiovas Surg 1985;89:115–120.
6. Astrom KJ. Theory and application of adaptive control— a survey. Automatica 1983;19:471–486.
7. Astrom KJ, Wittenmark B. Adaptive Control 2nd ed. New York: Addison-Wesley; 1994.
8. Goodwin GC, Sin KS. Adaptive Filtering, Prediction, and Control. Englewood Cliffs (NJ): Prentice Hall; 1984.
9. Ljung L. System Identification: Theory for the User. 2nd ed. Englewood Cliffs (NJ): Prentice Hall; 1998.
10. Arnsparger JM, McInnis BC, Glover Jr JR, Norman NA. Adaptive control of blood pressure. IEEE Trans Biomed Eng 1983;BME-30:168–176.
11. Meline LJ, Westenskow DR, Pace NL, Bodily MN. Computer controlled regulation of sodium nitroprusside infusion. Anesth Analog 1985;64:38–42.
12. Waller JL, Roth JV. Computer-controlled regulation of sodium nitroprusside infusion in human subjects. Anesthesiology 1985;63:A192.
13. Mansour NE, Linkens DA. Pole-assignment self-tuning control of blood pressure in postoperative patients: a simulation study. Proc Inst Electr Eng 1989;136, (Pt. D):1–11.
14. Voss GI, Katona PG, Chizeck HJ. Adaptive multivariable drug delivery: control of arterial pressure and cardiac output in anesthetized dogs. IEEE Trans Biomed Eng 1987;BME-34:617–623.
15. Kaufman H, Roy R, Xu X. Model reference control of drug infusion rate. Automatica 1984;20:205–209.
16. Pajunen GA, Steinmetz M, Shankar R. Model reference adaptive control with constraints for postoperative blood pressure management. IEEE Trans Biomed Eng 1990;BME-37:679–687.
17. Polycarpou MM, Conway JY. Indirect adaptive nonlinear control of drug delivery systems. IEEE Trans Auto Control 1998;AC-43:849–856.
18. Lainiotis DG. Partition: a unifying framework for adaptive systems II: control. Proc IEEE 1976;64:1182–1198.
19. He WG, Kaufman H, Roy R. Multiple model adaptive control procedure for blood pressure control. IEEE Trans Biomed Eng 1986;BME-33:10–19.
20. Martin JF, Schneider AM, Smith NT. Multiple-model adaptive control of blood pressure using sodium nitroprusside. IEEE Trans Biomed Eng 1987;BME-34:603–611.
21. Martin JF, Schneider AM, Quinn ML, Smith NT. Improved safety and efficacy in adaptive control of arterial blood pressure through the use of a supervisor. IEEE Trans Biomed Eng 1992;BME-39:381–388.
22. Martin JF, Smith NT, Quinn ML, Schneider AM. Supervisory adaptive control of arterial blood pressure during cardiac surgery. IEEE Trans Biomed Eng 1992;BME-39:389–393.
23. Yu C, Roy RJ, Kaufman H, Bequette BW. Multiple-model adaptive predictive control of mean arterial pressure and cardiac output. IEEE Trans Biomed Eng 1992;BME-39:765–778.
24. Garcia CE, Prett DM, Morari M. Model predictive control: theory and practices—a survey. Automatica 1989;25:335–348.
25. Rao RR, Palerm CC, Aufderheide B, Bequette BW. Automated regulation of hemodynamic variables. IEEE Eng Med Biol 2001;20 (1):24–38.
26. Rao RR, Aufderheide B, Bequette BW. Experimental studies on multiple-model predictive control for automated regulation of hemodynamic variables. Trans Biomed Eng 2003;50 (3):277–288.
27. Zadeh LA. Fuzzy sets. Inform Contr 1965;8:338–353.
28. Tong RM. A control engineering review of fuzzy systems. Automatica 1977;13:559–569.
29. Sugeno M. An introductory survey of fuzzy control, Inform. Science 1985;36:59–83.
30. Isaka S, Sebald AV. An optimization approach for fuzzy controller design. IEEE Trans Sys, Man, Cyber 1992;SMC 22:1469–1473.

31. Ying H, McEachern M, Eddleman DW, Sheppard LC. Fuzzy control of mean arterial pressure in postsurgical patients with sodium nitroprusside infusion. IEEE Trans Biomed Eng 1992;BME-39:1060–1070.
32. Hunt KJ, Sbarbaro D, Zbikowski R, Gawthrop PJ. Neural networks for control systems—a survey. Automatica 1992;28: 1083–1112.
33. Chen CT, Lin WL, Kuo TS, Wang CY. Adaptive control of arterial blood pressure with a learning controller based on multilayer neural networks. IEEE Trans Biomed Eng 1997;BME-44:601–609.
34. Kashihara K, et al. Adaptive predictive control of arterial blood pressure based on a neural network during acute hypotension. Ann Biomed Eng 2004;32:1368–1383.

See also BIOFEEDBACK; BLOOD PRESSURE MEASUREMENT; DRUG INFUSION SYSTEMS; HEMODYNAMICS; PHYSIOLOGICAL SYSTEMS MODELING.

BLOOD RHEOLOGY

ROGER TRAN-SON-TAY
University of Florida
Gainesville, Florida

WEI SHYY
University of Michigan
Ann Arbor, Michigan

INTRODUCTION

Blood rheology has had broad impact in our understanding of diseases and in the development of medical technology. Rheology is the science dealing with the flow and deformation of matter. Therefore, it encompasses work in mechanical, chemical, and biomedical engineering. It plays a vital role not only in the design, manufacture, and testing of materials, but also in the health of the human body. Biorheology is therefore concerned with the description of the flow and deformation of biological substances. More specifically, hemorheology, or blood rheology, deals with the rheological behavior of blood, including plasma and cellular constituents.

Blood flow is known to be responsible for the delivery of oxygen to tissue and the removal of carbon dioxide. However, it also plays a pivotal role in the transport of substances (nutrients, metabolites, hormones, cells, etc.) involved not only in the maintenance of the body and its immune response, but also in diseases. For example, cancer cells are transported through blood as they spread from one tissue to another in a process known as metastasis.

The rheology of blood is altered in a number of pathological conditions. Sickle cell disease is a genetic disease producing abnormal hemoglobin causing red blood cells (RBCs) to become crescent shaped when they unload oxygen molecules or when the oxygen content of the blood is lower than normal. Under these conditions, the sickle hemoglobin aggregates and the RBCs become rigid, and consequently obstruct and/or damage the capillaries. Sickle cell disease is also known as sickle cell anemia because of the abnormally low oxygen-carrying capacity of the blood due to an insufficient number of RBCs and an abnormal hemoglobin. During a heart attack or stroke, there is a partial or complete occlusion of blood vessels due to the formation of a blood clot that alters blood flow. It is clear that many diseases and factors (artherosclerosis, hypertension, vasodilator agents, etc.) can compromise blood flow by occluding vessels or modifying their rheological properties. However, the study presented here will focus mainly on the rheology of blood.

It is important to recognize that the rheological properties of blood and its components, that is, blood cells, are important in the aptitude of blood to perform its functions correctly. The ability of a blood cell to flow into capillaries or migrate through tissues is governed, but its rheological properties. In addition, flow is expected to affects cells in two ways: (*1*) the fluid moving over or around the cell will exert mechanical stress on the cell, and (*2*) the motion of fluid will alter the concentration of chemical species in the immediate surrounding of the cell, leading to the mass transport of nutrients, waste products, drugs, hormones, and so an, to and from the cell. Finally, blood rheology can also have an indirect but critical role in our immune system and in diseases since a given applied stress, such as fluid shear, can generate a signal that can induce or modify cellular response.

Blood rheology is an extremely broad subject that cannot be covered in a single review. Therefore, the scope of the present article is to provide an understanding of blood rheology, and an appreciation of its contributions to the improvement of our understanding and assessment of diseases. The article also provides a review of the most common methods used to measure the rheological properties of blood and blood cells.

RHEOLOGICAL PROPERTIES OF BLOOD

Is blood a Newtonian fluid? A Newtonian fluid is a fluid that has a viscosity that is constant and independent of the properties of the flow. This simple question is not easily answered because blood is a complex fluid. It is a non-Newtonian fluid that behaves as a Newtonian fluid under certain conditions. For example, for a shear rate $> 100\,\mathrm{s}^{-1}$ and a vessel/tube diameter $> 500\,\mu\mathrm{m}$, blood behaves as a Newtonian fluid. Blood is composed of formed elements (red cells, white cells, platelets, etc.) suspended in plasma. Blood cells are viscoelastic particles (possess both viscous and elastic properties), whereas plasma is Newtonian. Therefore, depending on the characteristics of the flow and size of the vessel (extent of deformation), the size and properties of the blood cells may not play a major role in the flow characteristics of blood. The behavior of blood needs to be described as a function of the size of the vessel and the rate of flow. Because of that behavior, the concept of apparent and relative viscosities is introduced. The viscosity value of a non-Newtonian fluid depends on the experimental conditions and instrument used to perform the measurement. Therefore, that measured viscosity is called the apparent viscosity, μ_{app}. The relative apparent viscosity, μ_{rel}, is defined as

$$\mu_{\mathrm{rel}} = \frac{\mu_{\mathrm{app}}}{\mu_{\mathrm{p}}} \tag{1}$$

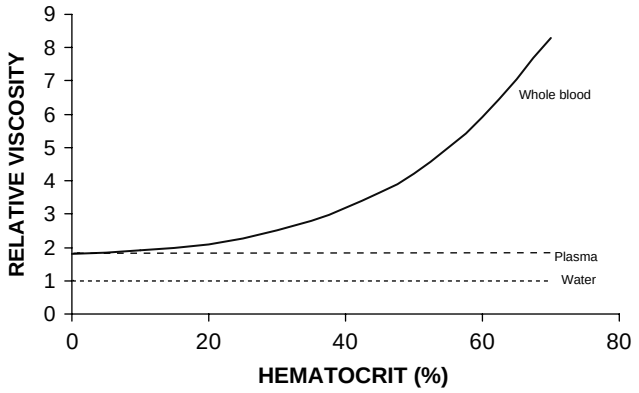


Figure 1. Effect of hematocrit on blood viscosity. Plasma has a relative viscosity of ~ 1.8 at 37 °C, that is a viscosity of ~ 1.2 mPa · s. Human blood at 40% hematocrit has a viscosity of ~3–4 mPa · s. However, it is a function of both hematocrit and shear rate.

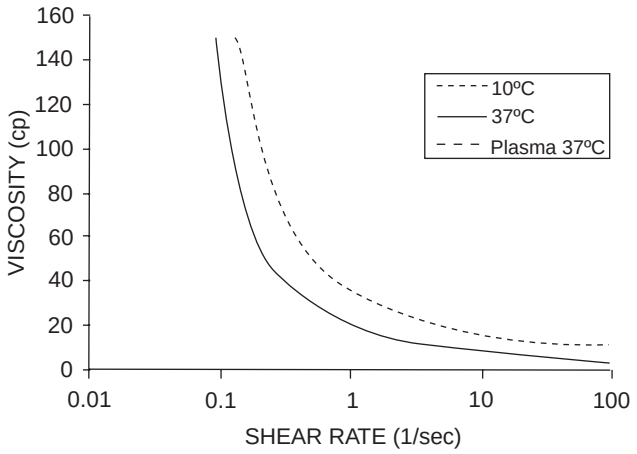


Figure 2. Effect of shear rate on blood viscosity. The Newtonian behavior of plasma and non Newtonian behavior of blood are clearly demonstrated. Plasma has a constant viscosity of ~ 1.2 mPa·s at 37 °C. The effect of temperature on blood viscosity is also shown.

where μ_p is the viscosity of plasma or other suspending medium.

In general, the viscosity of plasma is ~ 1.8 times the viscosity of water (termed relative viscosity) at 37 °C and is related to the protein composition of the plasma. Whole blood has a relative viscosity of ~ 4 depending on hematocrit (RBC concentration), temperature, and flow rate. The average hematocrit for a man and a woman is 42 and 38%, respectively. Hematocrit is an important determinant of the viscosity of blood. As hematocrit increases, there is a disproportionate (exponential) increase in viscosity (Fig. 1). For example, at a hematocrit of 40%, the relative viscosity is 4. At a hematocrit of 60%, the relative viscosity is ~ 8. Therefore, a 50% increase in hematocrit from a normal value increases blood viscosity by $\sim 100\%$. Such changes in hematocrit and blood viscosity occur in patients with polycythemia.

Because blood is non-Newtonian, the effect of shear rate is important. Figure 2 illustrates the shear thinning characteristic (decrease in viscosity as the shear rate increases) of blood at two different temperatures. On the other hand, it is clearly seen that plasma is Newtonian. It has a viscosity of ~ 1.2 cP or 1.2 mPa·s at 37 °C. The poise, P, is a unit of viscosity. The different viscosity units are related as follows: $1\,\mathrm{P} = 1\,\mathrm{dyn{\cdot}s{\cdot}cm^{-2}} = 0.1\,\mathrm{N{\cdot}s{\cdot}m^{-2}} = 0.1\,\mathrm{Pa{\cdot}s}$; therefore, $1\,\mathrm{cP} = 1\,\mathrm{mPa{\cdot}s}$. The fact that blood viscosity increases at low shear is one of the key factors for the initiation of atherosclerosis at specific sites in the arterial system. Increases in the viscosity of blood and plasma reflect clinical manifestations of atherothrombotic (formation of fibrinous clot) vascular disease. High blood viscosity invariably accompanies degenerative diseases. It is therefore not surprising that many treatments involve lowering blood viscosity to treat or prevent heart attacks, strokes, atherosclerosis, and so on.

Temperature also has a significant effect on viscosity. This can be seen in Figs. 3 and 4 where the effect of temperature on the viscosity of plasma and human blood is shown. Temperature has a similar effect on plasma and water. As temperature decreases, viscosity increases. Viscosity increases $\sim 2\%$ for each degree celcius decrease in temperature. This effect has several implications. For example, when whole-body hypothermia is used during certain surgical procedures, it increases blood viscosity and therefore augments resistance to blood flow.

The viscoelastic profile of normal human blood can be divided into three regions depending on the shear rate levels. In the low shear rate region ($\dot{\gamma} \leq 20\,\mathrm{s^{-1}}$), red cells are in large aggregates and as the shear rate increases, the size of the aggregates diminishes. Blood viscoelasticity is dominated by the aggregation properties of the red blood cells. In this region, human blood behaves like a Casson fluid with a small but finite yield stress (i.e., blood will not flow or deform unless the applied stress exceeds that critical stress),

$$\sqrt{\tau} = a + b\sqrt{\dot{\gamma}} \tag{2}$$

where a and b are constant (Fig. 5). The magnitude of the rheological parameters like yield stress and viscosity depends on various factors such as plasma protein concentration,

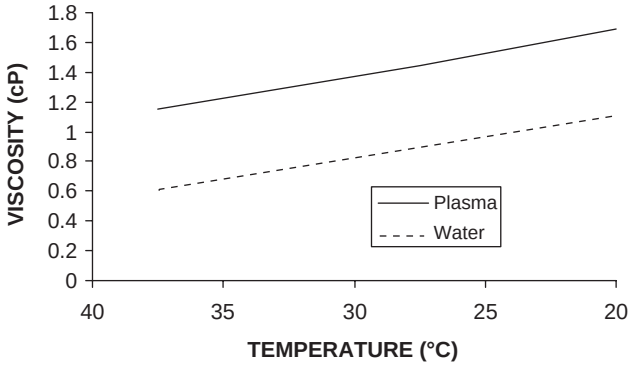


Figure 3. Effect of temperature on plasma viscosity. Temperature has similar effects on the viscosity of plasma and water. Viscosity increases $\sim 2\%$ for each degree celcius decrease in temperature.

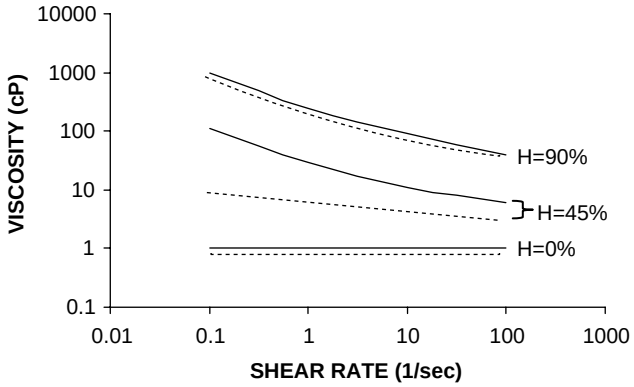


Figure 4. Effects of shear rate and hematocrit on blood viscosity. The shear thinning characteristics of blood is well illustrated in this figure. The solid and dashed lines represent, respectively, whole blood and washed red blood cells in a saline solution at 45 and 90% red cell volume concentrations. (Adapted from Ref. 1.)

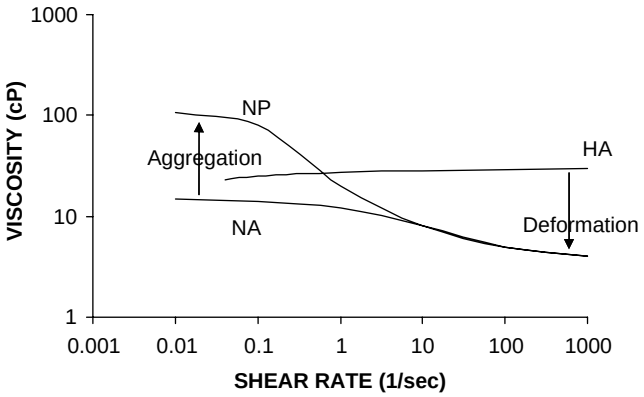


Figure 6. Effects of cell aggregation and deformability on blood viscosity. The logarithmic relation between the apparent viscosity and shear rate in three types of suspensions, each containing 45% human RBCs by volume is shown. Suspending plasma viscosity = 1.2 mPa.s; NP = normal RBCs in plasma; NA = normal RBCs in 11% albumin; HA = hardened RBCs in 11% albumin solution. (Adapted from Ref. 3.)

hematocrit, and properties of the blood cells. At low flow rates, there are increased cell-to-cell and protein-to-cell adhesive interactions that can cause erythrocytes (RBC) to adhere to one another and increase the blood viscosity. However, at shear rates $> 100\,s^{-1}$, cell aggregation and rouleaux formation break up and blood behaves as a Newtonian fluid with a viscosity of $\sim$3–4 mPa·s depending on the hematocrit and other factors (Fig. 6). In the mid-shear rate range ($20 \leq \dot{\gamma} \leq 100\,s^{-1}$), the cells are progressively disaggregated with increasing shear rate. Increasing shear rate causes the cells to deform and orient in the direction of flow, and the viscoelasticity of the blood is dominated by the deformability of the RBC. Figure 6 also demonstrates the effect of cell deformabilty on blood viscosity. It is seen that deformable cells lower the blood viscosity as compared to rigid ones.

However, when the dimensions of the cells are not negligible in comparison with the diameter of the vessel through which flow is occurring, the two phase nature of blood has to be recognized. Thus blood flow through vessels narrower than 500 μm in diameter is accompanied by several anomalous effects which can be directly traced to the two phase nature of blood. The important artifacts are the Fahraeus (a decrease in the hematocrit of the vessel–tube as compared to the larger feeding vessel–reservoir hematocrit) and Fahraeus–Lindqvist (a decrease in the vessel–tube apparent viscosity as compared to the larger feeding vessel–reservoir viscosity) effects. The effects are more pronounced as the vessel–tube diameter decreases. For vessel diameters < 10 μm (capillaries), blood cells must travel in single file and the flow must be analyzed as creeping (low Reynold number) flow of a Newtonian fluid with particles embedded in it. The transition from a single file to suspension flow occurs in the diameter range of 10–25 μm and this domain is difficult to analyze.

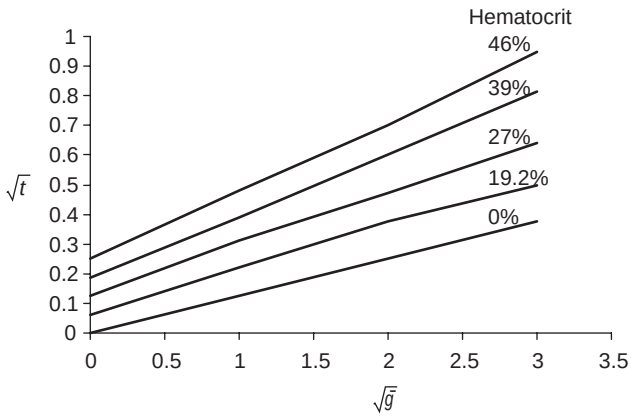


Figure 5. Blood behavior at low shear—casson behavior. These plots were generated from blood data obtained at 25 °C. They show that blood has a yield stress that depends on hematocrit. (Adapted from Ref. 2.)

To add to the complexity of blood behavior, vessels also affect blood flow characteristics. For example, the most noticeable feature of blood flow in the arteries is the pulsatile nature of the flow. However, this feature is lost in the microcirculation because its effect has been dampened by the viscoelastic blood vessels. The flow in the microcirculation occurs at very low Reynolds number (i.e., inertial forces due to transient and convective accelerations are negligible) and is determined by a balance of viscous stress and pressure gradient. Individual cells must be recognized. In the capillaries, 3–10 μm diameter vessels, cells flow in single line and their flow–deformation must be analyzed. Finally, in the veins, where $\sim$80% of the total volume of blood is located, the most noticeable feature is that vessels can collapse and that their mechanical properties cannot be neglected.

The above descriptions represent just some of the rheological characteristics of blood, but many more factors can affect its behavior. However, they prove the point that blood is an extremely complex fluid with many facets and that only a few have been unveiled so far.

CORRELATION BETWEEN BLOOD RHEOLOGICAL PROPERTIES AND CLINICAL CONDITIONS

Blood is a complex fluid whose flow (rheological) properties are significantly affected by the arrangement, orientation, and deformability of red blood cells. Variations in blood rheology among healthy individuals are very small. Thus, changes due to disease or surgical intervention can be readily identified, making blood rheology a useful clinical marker. Variations in blood rheology are observed in such conditions as cardiovascular disease, peripheral vascular disease, sickle cell anemia, diabetes, and stroke.

Although studies of blood rheology date from at least the early studies of Poiseuille (4), the discipline of clinical hemorheology is relatively new. It underwent a rapid growth in 1970–1980, in large part due to support by pharmaceutical companies and equipment manufacturers. Various instruments and devices were developed specifically for studying blood rheology.

Some of the early clinical tests dealing with blood rheology were on blood coagulation and the formation of blood clots. Most people think of blood in its liquid state, but its ability to thicken into a blood clot is a vital part of the body's natural defense. This process of forming a clot is referred to as coagulation. Blood coagulation, or blood clotting, is a complex process involving platelets, coagulation factors present in the blood and blood vessels. If blood becomes too thin, it loses the ability to form the blood clots that stop bleeding. When blood becomes too thick, the risk of blood clots developing within the blood vessels rises creating a potentially life-threatening condition. Blood disorders occur when hemostasis falls out of balance. Hemostasis is achieved when blood chemicals, hormones and proteins are correctly balanced. Hemostasis refers to the complicated chemical interplay that maintains blood fluidity (e.g., viscosity, elasticity, and other rheological properties).

Coagulation, or the lack thereof, is a key factor in various diseases. Sometimes thrombi (large clots) can completely occlude vessels. This can lead to ischemia, and ultimately death in any part of the body. Myocardial infarction and stroke are among the major life-threatening conditions caused by vessel occlusion due to clots. Conversely, there are various coagulatory disorders in which thrombus formation does not occur when it should. These bleeding disorders include various forms of hemophilia [e.g., (5,6)].

A great deal of research has focused on the effects of rheology on thrombus formation. Various *ex vivo* and *in vitro* systems have been designed to mimic *in vivo* blood flow in order to study thrombus formation within the circulatory system (7), and on various devices. For example, these systems have been used to model blood flow in order to study thrombus formation on stents (8) and mechanical heart valve prostheses (9). In addition, some research has focused on the effect of shear on thrombus dissolution. These studies suggest that thrombi lysis is accelerated with increasing shear rates (10,11).

It is impossible to cite all the contributions of blood rheology to our understanding of diseases in this review, but it is clear that viscosity was and still is clinically the most commonly used rheological property. The principal factors determining blood viscosity are hematocrit, plasma viscosity, cell aggregation, and cell deformability. Earlier rheological work was mainly performed on whole blood and on RBCs because the latter are by far the most numerous cells in our body (99% of the blood cells are RBCs). However, in the last two decades, the focus has shifted toward understanding the rheology of leukocytes or white blood cells (WBC) because they have been found to be bigger and more rigid that RBC. The major motivation behind all these blood cell studies is that the ability of a cell to deform and flow through the capillaries and/or to migrate in the tissue is determined by its rheological properties, and this ability is vital in its response to disease/infection. These properties, in turn, are a manifestation of the underlying structure of the cell and the organization of the structural components (microfilaments (F-actin), microtubules, intermediate filaments, lipid bilayer) in the cellular cytoplasm and cortex.

Because blood rheology is a very broad subject, this article focuses on the role of RBC deformability in clinical studies. The role of other blood cell types, cytoskeleton, proteins, adhesion molecules, and so on., although important and of interest, is beyond the scope of this article and will not be addressed.

RBC Deformability

Deformability is a term used to describe the ability of a body (cell in the present context) to change its shape in response to an applied force. A very important characteristics of a normal RBC is that it has a surface area $\sim 30\%$–40% greater than that of a sphere of equal volume. Other major determinants of RBC deformability include rheological properties of the cell membrane, and intracellular fluid.

Cell deformability can be determined by direct microscopic measurement (micropipette) or indirect estimation (filtration). By using micropipettes with diameters $\geq 3\,\mu m$, the entire RBC can be aspirated. The deformability of the cell can be estimated from the pressure required for its total aspiration.

The importance of cell deformability is well established in the studies of the rheological behavior of RBCs in the capillary network. It was clearly demonstrated that reductions in RBC deformability may adversely affect capillary perfusion (12) and that many diseases manifest reductions in RBC deformability (13–15). Of the many determinants of capillary perfusion, the size of the undeformed RBC relative to the capillary diameter may play the greatest role in affecting capillary perfusion. For example, studies of the passage of RBCs through capillary-sized pores of polycarbonate sieves (16) reveal that the flow resistance may increase 30–40 times as the ratio of pore to cell diameter is reduced from 1 to 0.1. Furthermore, after entry into a capillary, the ability of RBCs to deform may play an equally important role as RBCs negotiate irregularities in the capillary lumen, as manifested by encroachment of endothelial cell nuclei on the capillary lumen (17). It was also demonstrated that the microvascular network may passively compensate for increased RBC stiffness by

shunting RBCs within the capillary network through pathways of lesser resistance (18).

There are many methods for assessing the erythrocyte deformability but only two (filtration of RBCs through pores of 3–5 mm diameter and the measurement of RBC elongation using laser diffractometry) have been widely applied clinically. A brief description of these two methods is provided below.

Erythrocyte Filtration. Filtration method has been commonly used to study the deformability of RBC. The basic idea is to force the RBC suspension to flow through 3–5 μm pores (by using a negative or positive pressure or gravity), and obtain the relationship between pressure and flow rate to estimate the deformability of the cells. Either the flow rate is measured under a constant pressure or the pressure is measured under a constant flow rate. Contaminants, such as WBC, which is poorly deformable affect the experiment by plugging the pores.

The techniques for whole-blood filtration are almost all derived from that described by Reid et al. in 1976 (19). The results of these methods are expressed as volume of blood cells (VBCs) in the time unit. However, this technique is susceptible to aggregation of RBCs and contamination with leukocytes. A modified version of the apparatus was developed to reduce these problems (20). Nevertheless, WBC contamination remains an issue with whole-blood filtration techniques.

A common draw back among all these filtrometry-based instruments is the lack of any measure of individual cell volume, thereby making it difficult to distinguish changes in RBC filtration due to the volume distribution (or aggregates) within the RBC sample from those due to intrinsically less deformable cells.

Another filtration technique that is commonly used is the Bowden assay (21,22). However, this assay involves the migration of cells (WBCs) through a filter membrane with pores of defined diameter and is beyond the scope of this article.

Erythrocyte Elongation. The Ektacytometer (23) combines viscosity with laser diffractometry. It consists of a transparent cylindrical Couette or a cone-plate viscometer, which allows a helium–neon laser beam to pass through the erythrocyte test suspension during rotational shear. The laser diffracted image becomes elliptical as the RBCs are sheared, and the ratio of the major over minor axes of the image is called the elongation index. This dynamic measurement of RBC elongation has been used for the rheological studies of congenital defects of RBC membrane protein (24), and many blood disorders (25–27).

It is important to remember that the two methods described above provide information on the bulk deformability of the RBC only, and are not suited for characterizing the deformability of subpopulations. Alternative methods need to be used for these studies. Some of the methods that have been developed for specifically characterizing the rheological properties of individual cells are provided in the next section.

RHEOLOGICAL PROPERTIES MEASUREMENTS

The goal of this section is to provide an overview of the most commonly used techniques for characterizing the rheological properties of blood. Therefore, devices will be divided into two groups: one for characterizing fluids, the other for individual cells.

Techniques for Measuring the Rheological Properties of a Fluid

Cylindrical Tube. The first studies of blood rheology have been done in cylindrical tubes. In his quest toward developing a better method for measuring blood pressure, French physician and physiologist Jean Louis (or sometimes called Leonard) Marie Poiseuille (1799–1869) studied the flow of liquid through tubes. (There is some confusion about Poiseuille's precise name and year of birth, sometimes quoted as 1797.) In 1838, he established a series of meticulously executed experiments: At a given temperature the rate of water flow through tubes of very fine bore is inversely proportional to the length of the tube and directly proportional to the pressure gradient and to the fourth power of the tube diameter. In 1840 and 1846, he formulated and published an equation known as Poiseuille's law (or Hagen-Poiseuille law, named also after the German hydraulic engineer Gotthilf Heinrich Ludwig Hagen who independently carried out friction experiments in low speed pipe flow in 1840) based on his experimental pipe flow observation. Little is known of the life of Jean Leonard Marie Poiseuille. However, he made important contributions to the experimental study of circulatory dynamics. His law can be successfully applied to blood flow in capillaries and veins, and to air flow in lung alveoli, as well as for the flow through hypodermic needle or tubes, in general (28,29).

The derivation of Poiseuille's law for a Newtonian fluid, that is, the viscosity of the fluid is constant and independent of the properties of the flow. (Viscosity is a property of fluid related to the internal friction of adjacent fluid layers sliding past one another, as well as the friction generated between the fluid and the wall of the vessel. This internal friction contributes to the resistance to flow.) For example, water and plasma are Newtonian fluids. For a Newtonian, laminar (nonturbulent) case, the flow through a cylindrical tube is one-dimensional (1D) (Fig. 7) reaching the fully

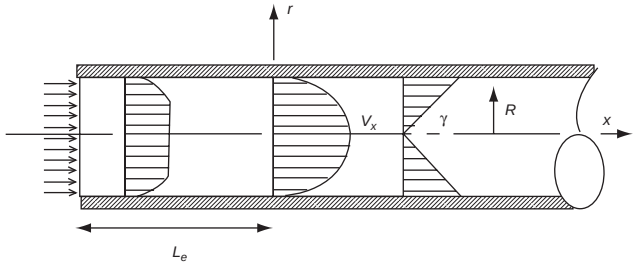


Figure 7. Flow in a cylindrical tube. Fully developed, laminar, viscous flow in tubes produces a parabolic velocity profile. The shear stress varies linearly with the radial distance r. Parameters are as follows: velocity field, $V_x = V_x(r)$, $V_r = 0$, $V_\theta = 0$; shear rate, $\dot{\gamma} = -(\partial V_z/\partial r)$; lines of shear, straight lines parallel to the tube axis; shearing surfaces, concentric cylinders.

developed state, that is, exhibiting no variation in velocity profiles in the streamwise, x direction, and the streamwise velocity, V_x, has the following distribution on any cross-section

$$V_x = \frac{R^2 \Delta P}{4 \mu L}\left[1 - \left(\frac{r}{R}\right)^2\right] \tag{3}$$

where ΔP is the pressure drop between two points located at a distance L apart, R is the tube radius, r is the radial distance from the tube axis, and μ is the fluid viscosity.

For 1D laminar flows in pipe, the pressure gradient, ΔP, necessary to produce a given flow rate, Q, is proportional to the viscosity and, as shown from the above equation, inversely proportional to the fourth power of the tube radius,

$$\frac{\Delta P}{L} = \frac{8 \mu Q}{\pi R^4} \tag{4}$$

This equation has important clinical implications. It tells us that for a given pressure drop, a 10% change in vessel radius will cause $\sim$50% change in blood flow. Conversely, for a fixed flow, a 10% decrease in vessel radius will cause $\sim$50% increase in the required pressure difference. Poiseuille law tells us the consequences of having a reduced vessel lumen like in arteriosclerosis.

Thus, the average fluid velocity can be expressed in terms of the volumetric flow rate Q or pressure ΔP

$$\overline{V} = \frac{Q}{\pi R^2} = \frac{R^2 \Delta P}{8 \mu L} \tag{5}$$

For a Newtonian fluid, the shear stress distribution in the tube is linear with r,

$$\tau = \mu \dot{\gamma} = -\mu \frac{dV_z}{dr} = \frac{r \Delta P}{2L} \tag{6}$$

The shear stress is the frictional force per unit area as one layer of fluid slides past an adjacent layer. Therefore, the maximum shear stress occurs along the tube wall and is equal to

$$\tau_z = \frac{R \Delta P}{2L} = \frac{4 \mu Q}{\pi R^3} = \frac{4 \mu \overline{V}}{R} \tag{7}$$

The viscosity of blood and other fluids have been characterized with cylindrical tube devices. For example, Cannon–Fenske viscometers are cylindrical tubes used to measure the viscosity of fluids. However, they are not commonly used for measuring non–Newtonian fluids because the shear rate generated in these devices is not constant so its effect on viscosity cannot be easily characterized.

Viscometers. Viscometers are designed to measure the viscosity of fluids. They come in many forms (e.g., concentric cylinders, parallel disks) (30), but the review will focus only on instruments that have been used to characterize biological fluids (31,32). The most common is the cone-plate arrangement (Fig. 8) because it produces a linear velocity profile and, consequently, a constant shear rate throughout the gap for small cone angles.

For a cone-plate viscometer of radius R and cone angle α_o, the relevant parameters, such as viscosity and shear rate, can be found by setting the angular speed of rotation

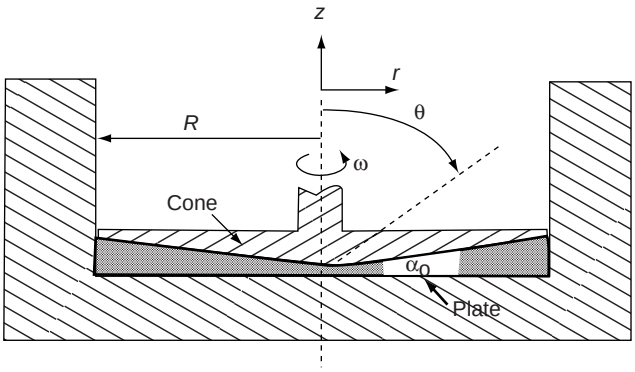


Figure 8. Cone-plate viscometer. For small cone angles, the flow between the cone and the plate is linear. Parameters are as follows: velocity field, $V_\phi = V_\phi(r,\theta)$, $V_\theta = 0$, $V_r = 0$; shear rate, $\dot{\gamma} = -(\frac{1}{r})(\frac{\partial V_\phi}{\partial \theta})$; lines of shear, circles of constant r and z; shearing surfaces, cones of constant θ.

of the cone, ω, and observing the resultant torque, T. These relationships are provided through the expressions of the shear stress, τ, and shear rate, $\dot{\gamma}$.

$$\tau = \frac{3T}{2\pi R^3} = \frac{\mu \omega}{\alpha_o} \tag{8}$$

$$\dot{\gamma} = -\frac{\sin\theta}{r}\frac{d}{d\theta}\left(\frac{V_\phi}{\sin\theta}\right) \cong -\frac{1}{r}\frac{dV_\phi}{d\theta} = \frac{r\omega}{d} = \frac{\omega}{\alpha_o} \tag{9}$$

In Eq. 9, d represents the gap width at a radial distance, r.

In order to measure the elastic properties, dynamic testing needs to be performed. These viscometers have to be run in an oscillatory mode so that elastic effects can be detected. In general, the rheological properties of the fluid can be described in terms of the complex viscosity η^*, or complex modulus G^*. These complex parameters are composed of a viscous and an elastic components, and are related to each other by the angular frequency of the oscillation ω ($G^* = \omega\eta^*$). It is common to mix the notation and use the viscous component, η', of η^*, and the elastic component or storage modulus, G', of G^*, to characterize the viscoelastic properties of the fluid. The viscous and elastic components represent, respectively, energy lost irreversibly and stored reversibly by the sample during an oscillatory cycle.

Microrheometers. As opposed to viscometers, rheometers are devices that measure not only viscosity, but also other rheological properties, like elasticity and yield stress. However, that characterization is very casual since many viscometers have been modified, as described above, to measure the viscoelastic properties of fluids.

As their names indicate, microrheometers have been developed to measure the rheological properties of small volume biological fluids. Typically, these machines require one drop of fluid or less. The design of the magneto-acoustic ball microrheometer for measuring the rheological properties of a liquid is shown in Fig. 2 (33). This instrument requires a much smaller sample size (20 μL, i.e., about a drop) than traditional rheometers, and opaque suspensions

can be studied with it. The small-volume rheometer permits accurate temperature control and rapid temperature changes for kinetics studies, if needed (34).

The instrument itself consists of a magnetically driven 0.8 or 1.3 mm stainless steel ball that is tracked by ultrasonic echo location as it moves within the sample fluid. Using a system consisting of a time-to-voltage Converter (TVC), a pulse generator, a differential amplifier, and an oscilloscope (150 MHz), ball displacements as small as 3 μm are measured. The improved microrheometer (34) is capable of accurately measuring the viscosity of water in the very short chamber. Two measurements can be made (*1*) a falling-ball viscosity and (*2*) an oscillating-ball frequency dependent viscoelastic measurement. Parameters measured are η, the falling ball or steady-state viscosity; η', the viscous or loss modulus; and G', the elastic or storage modulus.

An experiment with the falling ball consists of dropping the ball along the centerline of a 10 mm long tube with a radius of 1.6 mm. The tube is surrounded by a large flow-through chamber for accurate temperature control (Fig. 9). The terminal velocity of the ball, V, is inversely proportional to the viscosity, η. In rheology, it is common to denote the viscosity as η for a viscoelastic fluid. This velocity–viscosity relationship is readily derived from the Stokes drag equation:

$$\eta = 2[(\rho_s - \rho)R^2 g]/9\,VK \tag{10}$$

where ρ_s and ρ are the ball and fluid density, respectively, R is the ball radius, g is the acceleration due to gravity, and K is the wall correction factor to account for the tube wall effect.

Oscillating ball experiments are operated over a frequency range of 1–20 Hz, whereby the sinusoidal driving force and the resulting sinusoidal sphere displacement are recorded. The magnitude of the displacement sinusoid and its phase shift relative to the driving force provide a measure for η' and G'. The system is calibrated with a series of Newtonian silicone oils from 0.1 to 100 P (1 P = 0.1 Pa·s). For a ball oscillating in a viscoelastic medium, the viscous and elastic moduli, when inertia is negligible, are defined as

$$\eta' = \frac{F_o \sin\phi}{6\pi K R \omega X_o} \tag{11}$$

and

$$G' = \frac{F_o \cos\phi}{6\pi K R X_o} \tag{12}$$

where F_o is the magnitude of the oscillating magnetic force, ω is the angular frequency of oscillation ($\omega = 2\pi f$), and X_o is the amplitude of the ball displacement.

The viscoelastic properties of blood have been characterized using the instruments described above in an oscillatory mode. However, these viscoelastic data, although useful as a tool for comparing different blood types and diseases, are not widely used because they are difficult to relate to the mechanical properties of the blood cells. For a non Newtonian fluid, the apparent viscosity (i.e., the slope of the curve of shear stress vs. shear rate at a particular value of shear rate) is used instead of viscosity since the latter is no longer a constant value and will depend on the rate and extent of deformation.

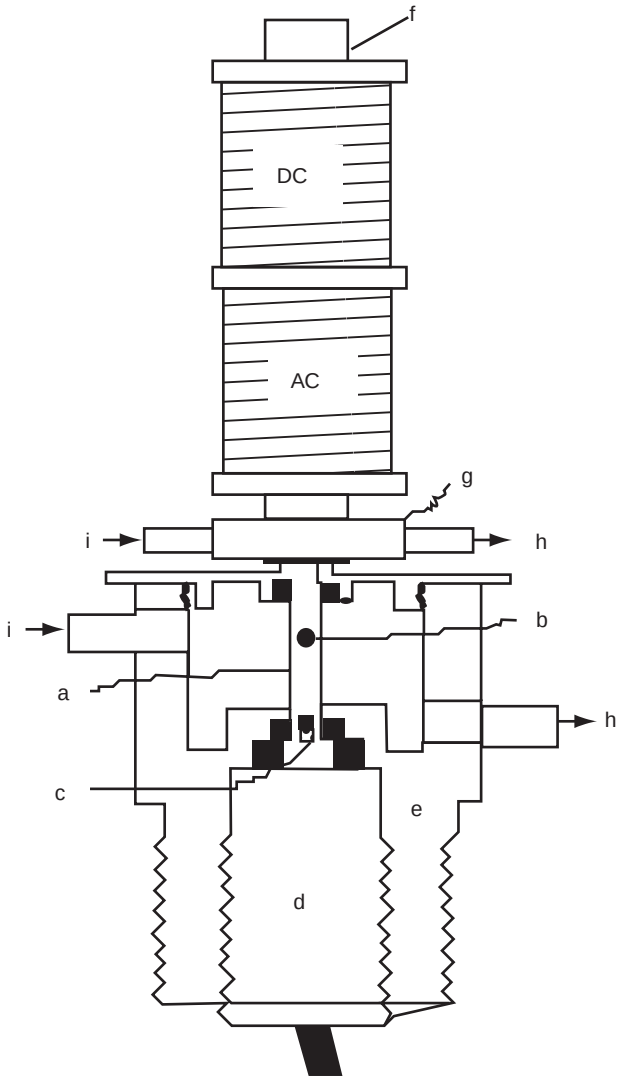


Figure 9. The magneto-acoustic ball microrheometer. (a) sample chamber, (b) stainless steel ball, (c) ultrasound crystal, (d) ultrasound transducer, (e) water jacket, (f) electromagnet, (g) electromagnet bath cap, (h) water flow outlet, (i) water flow inlet.

Techniques for Measuring the Rheological Properties of Blood Cells

Micropipette. The most popular technique for measuring the mechanical properties of blood cells is the micropipette technique (35). The micropipette manipulation technique has been used for studying liquid drops, cells, and aggregates. It has been used to investigate the effects of diseases (36,37) as well as treatments (38,39).

Micropipettes are made from 1mm capillary-glass tubing pulled to a fine point by quick fracture to give an orifice of desired diameter with a square end. The micropipette technique has been extensively used to characterize the mechanical properties of the RBC, but its use is limited in

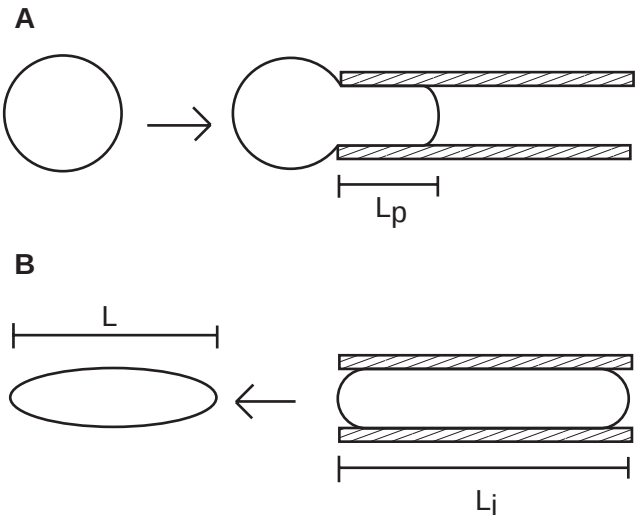


Figure 10. Schematic of the micropipette technique. Two micropipette experiments for determining cell viscosity and surface tension are depicted. (a) Aspiration experiment; for a given aspiration pressure, the length of the aspirated cell, Lp, is tracked as a function of time. (b) Recovery experiment; a cell fully aspirated inside a pipette is expelled from it. The length of the cell, L, is recorded as a function of time.

practice by the complexity of the theory associated with the biconcave shape of the cell. As an example of the micropipette technique, work on the characterization of WBCs will be presented below. The theoretical work is simplified because the shape of a WBC can be treated as a sphere. An equivalent simulation for a RBC will require extensive numerical work. Two typical types of experiment, aspiration and recovery (Fig. 10), are usually performed to determine the mechanical properties of individual WBCs (40–43).

Aspiration. Passive leukocytes (WBCs) are aspirated at a constant pressure into a micropipette. The length of the aspirated cell, L_p, is measured over time to generate an aspiration curve (Fig. 10a). Viscosity values, μ, can be derived from the slope, dL_p/dt, of the aspiration curves (42):

$$\mu = \frac{(\Delta P)R_p}{(dL_p/dt)m(1-\frac{1}{\bar{R}})} \tag{13}$$

where ΔP is the aspiration pressure, R_p is the pipet radius, $\bar{R} = R/R_p$, R is the radius of the cell outside the pipette, and $m = 6$. Figure 11 shows the aspiration of a white blood cell (lymphocyte) into a 4 μm diameter pipette. Fluorescence is used to better see the deformation of the cell nucleus.

Recovery. White blood cells are drawn by a small suction pressure into a micropipette, held there for ~15 s, and quickly expelled out. The changing length of the cell, L, as it recovers its spherical shape (Fig. 10b), is recorded as a function of time, t, and is described by a polynomial (43):

$$\frac{L}{D_o} = \frac{L_i}{D_o} + A\bar{t} + B(\bar{t})^2 + C(\bar{t})^3 \tag{14}$$

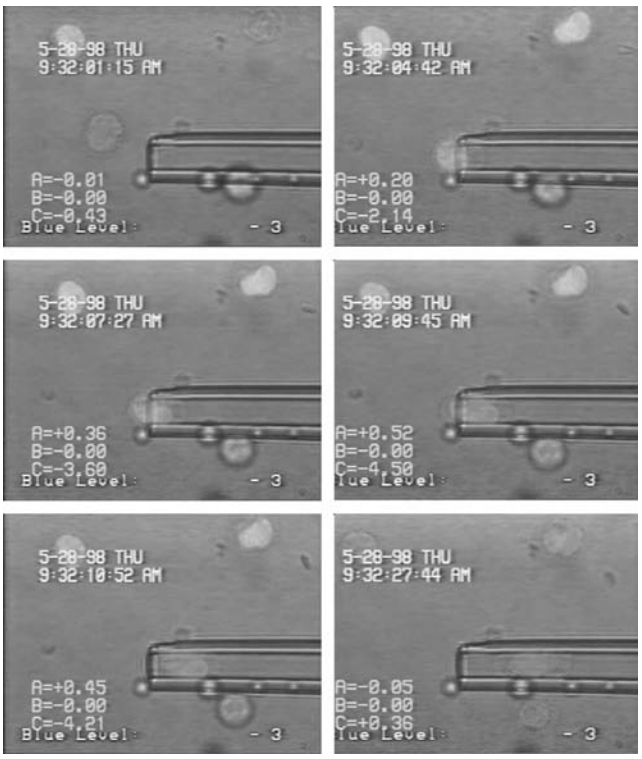


Figure 11. Flow of a lymphocyte inside a 4 μm pipette. The flow of an 8 μm lymphocyte (a WBC) inside a 4 μm diameter micropipette as a function of time is shown. Fluorescent technique was used to track the cell nucleus as well as the cellular membrane.

where A, B, C are known functions of (L_i/D_o). The parameters L_i and D_o are the initial deformed length and resting diameter of the cell, respectively. The variable $\bar{t} = 2t/[(\mu/T_o)D_o]$ represents a dimensionless time, where μ is the cell viscosity, and T_o is the surface tension of the membrane.

Figure 12 shows a lymphocyte (about 8 μm in diameter) aspirated inside a 4 μm diameter pipette (top picture). The cell is then expelled from the pipette and recovers its initial shape (bottom, left-hand side pictures). Pictures generated on the right hand side are from numerical simulation (44). In addition to the experimental techniques, significant progress has been made in the computational capabilities to simulate the dynamic behavior of blood at both large vessel and cellular scales (45).

Rheoscope

To allow direct observation of suspended cells during shear stress application, a modified cone-plate viscometer, called a rheoscope (Fig. 13), has been developed (46). In which the cone and plate counterrotate. This gives the advantage that a particle midway between the cone and plate is subjected to a well-defined shear stress field and remains nearly stationary in the laboratory frame of reference so that it can be studied without the help of high speed cinematography. It is important to note that, for an identical speed of rotation, that the shear rate generated in the rheoscope is twice that in the

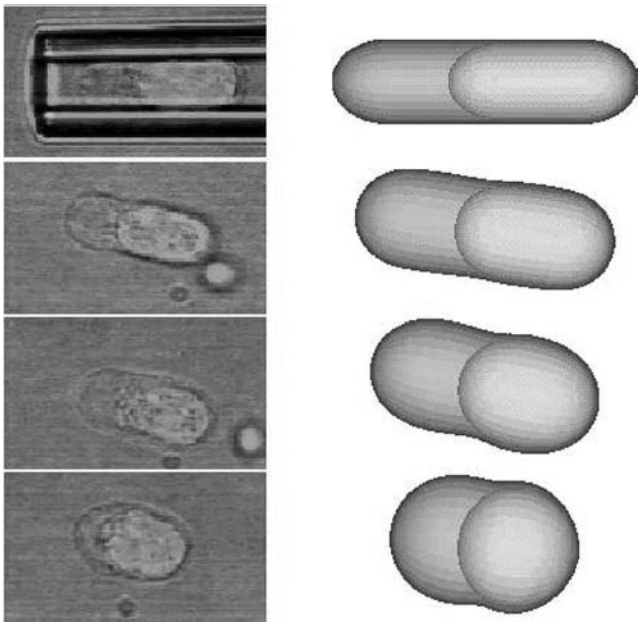


Figure 12. Pictures of a lymphocyte inside a pipette and its recovery. The top left frame is a picture of a lymphocyte (a type of leukocyte) aspirated inside a micropipette, the frames below it show the cell recovering its initial shape. Pictures on the right hand side are those generated by a theoretical compound drop model (44).

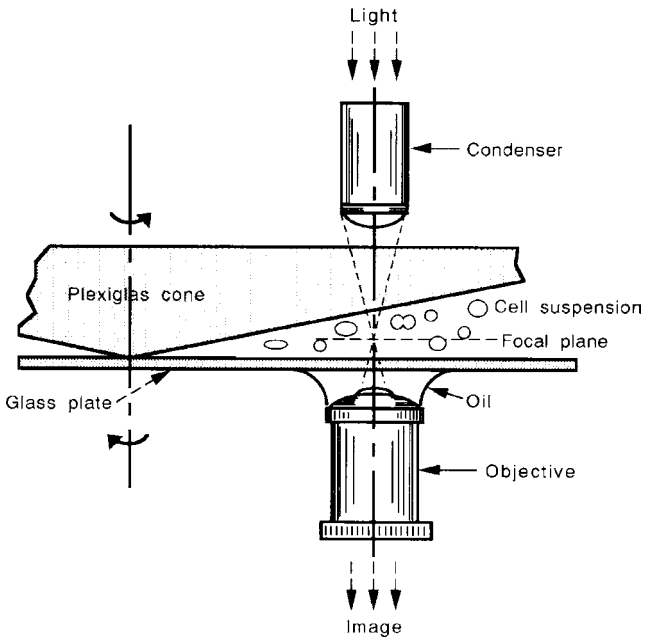


Figure 13. The Rheoscope. The main feature of this instrument is the counterrotation of the cone and plate. This gives the advantage that a particle midway between the cone and plate, subjected to a well-defined shear stress can be studied without the help of high speed cinematography. At a distance r from the axis of rotation, the shear rate is equal to $\dot{\gamma} = 2r\omega/d$, where d is the local gap width and ω is the angular velocity of the cone and plate

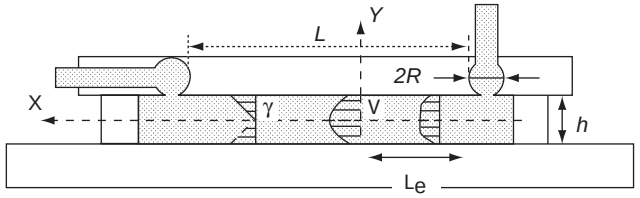


Figure 14. Parallel-Plate Flow Channel. The velocity profile is parabolic, and the shear rate is linear with y. Parameters are as follows: velocity field, $V_x = V_x\ (y)$, $V_y = 0$, $V_z = 0$; shear rate, $\dot{\gamma} = (\partial V_x/\partial y)$; lines of shear, straight lines parallel to the channel axis; shearing surfaces, plane surfaces parallel to the channel axis.

conventional cone-plate viscometer because of the counter rotation of the cone and plate. Effects of shear and mechanical properties of individual cells and anchorage dependent cells have been determined with the rheoscope (e.g., 1,36,47–49). The rheoscope is popular for studying RBC under shear because the analysis of the behavior of RBC is simplified since its shape is an ellipsoid.

Parallel Plate Flow Channel. Another popular technique for characterizing the rheological properties of blood cells under flow or for studying the effects of shear stress on anchorage dependent cells is the parallel plate flow system (Fig. 14).

The flow between infinite parallel plates is often referred to as plane Poiseuille flow. The velocity field reduces to one component in the direction of flow, V_x,

$$V_x = \frac{h^2 \Delta P}{8\mu L}\left[1 - \left(\frac{2y}{h}\right)^2\right] \tag{15}$$

where h is the distance between the plates, μ is the fluid viscosity, ΔP is the pressure drop between the inlet and outlet located at a distance L apart, and y is the vertical distance from the origin taken at the centerline of the channel.

The relationship between the pressure drop and volumetric flow rate Q is

$$\Delta P = \frac{12\mu QL}{wh^3} \tag{16}$$

where w is the channel width.

From the velocity field, Eq. 5, the shear rate across the channel gap is readily derived:

$$\dot{\gamma} = -\frac{dV_x}{dy} = \frac{12Q}{wh^3}y \tag{17}$$

For a Newtonian fluid, the relationship between shear rate and shear stress, τ, is linear, and the shear stress across the flow channel gap is

$$\tau = \mu\dot{\gamma} = \frac{12\mu Q}{wh^3}y \tag{18}$$

From the above equation, the shear stress is zero along the channel centerline, and the maximum shear stress, τ_s, is at the surface of the plate

$$\tau_s = \frac{h\Delta P}{2L} = \frac{6\mu Q}{wh^2} \tag{19}$$

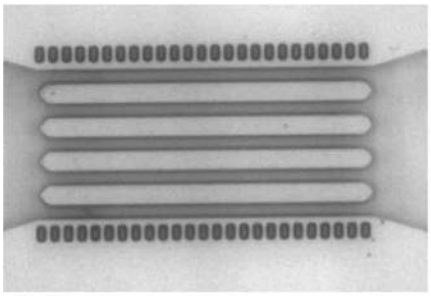
Vote stressed that, although they allow direct observation of individual particles, these devices do not provide a direct measurement of the particle viscosity. It is necessary to know the material constitutive properties of the particles, that is, expressions that relate stresses to strains, or to develop a mathematical model in order to determine the mechanical properties of the particles. Existing models can describe fairly accurately the rheological behavior of blood cells, but the exact rheological property values of these cells, with the exemption of red blood cells, are not known. That topic will not be covered here. For a review on the mechanical properties of blood cells, the reader is referred to Waugh and Hochmuth (35), and for a discussion on the discrepancy between the rheological data on white blood cells reported in the literature to Kan et al. (44).

CONCLUSION

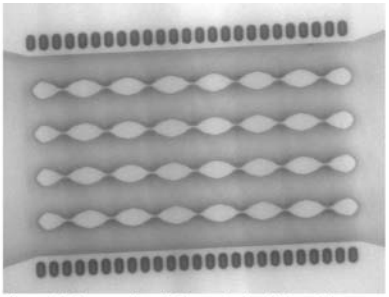
Some of the major advances in blood rheology are now being linked to the development and application of microfabrication to medicine. As reviewed by Voldman et al. (50) and Shyy et al. (51), these new tools will be used to better characterize the rheological properties of blood and blood cells, as well as to detect and diagnose cardiovascular and blood related diseases. Microfabrication is a process used to construct objects with dimensions in the micrometer to millimeter range. These objects are composed of miniature structures that can include moving parts such as cantilevers.

Soft lithography is a tool for micro/nanofabrication. It provides a convenient and effective method for the formation and manufacturing of micro- and nanostructures. Soft lithography is the collective name for a set of techniques that include replica molding, microcontact printing, microtransfer molding (52). The major advantages of soft lithography are that it is very fast as compared to conventional methods, relatively inexpensive, and applicable to almost all polymers. It is possible to go from design to production of replicated structures in < 24 h. In soft lithography, a master mold is first made by a lithographic technique, and an elastomeric stamp is then cast using the master mold. The elastomeric stamp with patterned relief structures on its surface is used to generate patterns and structures with feature sizes as small as 30 nm (53). Polydimethylsiloxane (PDMS) is the polymer of choice for many biological applications because it is optically transparent, isotropic, homogeneous, durable, and has interfacial properties that are easy to modify (53,54). As opposed to photolithography, soft lithography provides a mean for producing nonplanar surfaces.

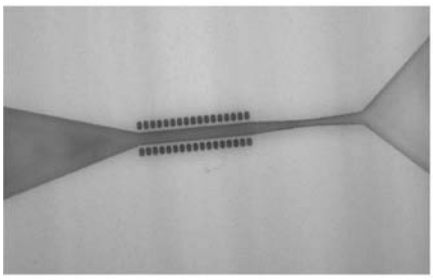
Microfabricated devices, also known as microelectromechanical systems (MEMS), are ideal tools for studying specific biological phenomena, for example, cell adhesion, as well as biological systems such as the microcirculation. For example, the fabrication of in vitro blood vessels can help to (*1*) determine blood cell distributions during blood flows through both arterial and venous type bifurcations, with successive bifurcations arranged as in microcirculation; (*2*) validate computer simulations and experimental methods used for *in vivo* measurements of parameters such as blood average velocity and hematocrits; and (*3*) to separate vessel or wall effects from hemodynamics effects. Figure 15 shows channels that were created using soft lithography. The different configurations shown are a series of 10 mm diameter channel in parallel, a series of



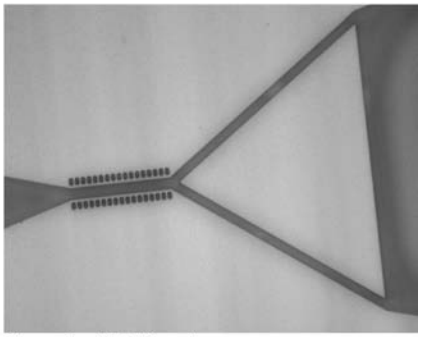
Series of 5 channels (150 μm long × 10 μm width) in parallel with smooth surface



Series of channels with endothelial cell pattern surface



Channel with constriction, smooth surface



Channel with bifurcation
Smooth surface

Figure 15. Photographs of microfabricated channels. The dashed lines on the top and bottom of the channels are length scales and each line represents 100 μm. Channels were created with inner diameters ranging from 5 to 100 μm.

channels with an endothelial cell like pattern surface, a channel with a constriction, and a channel with a bifurcation.

The application of MEMS in the area of microfluidics is a new and emerging field. It includes the development of miniature devices in fluid control, fluid measurement, and medical testing. Because of our need and interest in understanding, preventing, and treating diseases, most of the emerging MEMS applications are expected to be in biomicrofluidics. One of these applications is the development of gene chips and related deoxyribonucleic acid (DNA) tools. Others applications include microscale chemical analysis, known as microelectrophoresis, blood chemistry measurements using micromachined thin-film sensor, micropumps, drug delivery systems, glucose sensors, and chip-based microflow cytometry devices.

The major advantages of MEMS-based devices are that they are smaller and have the potential to be less expensive, more durable, and more reliable than conventional techniques. In addition, they can perform chemical tests much faster.

Finally, another area that will benefit from the advancement of technology is the development of blood substitutes. This is a needed area since the demand for blood continues to outpace the supply, especially in developing countries. With new technologies, the life span of blood substitutes based on cell-free hemoglobin, which is presently too short, could be lengthened. New methods could also be developed in order to produce human red cells, and other cell types, in culture.

Needless to say that, in order to be successful, all these new technical advances will need to be combined with an improved understanding of cell biology and rheology. This advancement will also depend on the successful incorporation of more sophisticated mathematical and computational techniques. In general, the field of biorheology is moving towards better understanding phenomena at the molecular level. For example, the knowledge of the signal transduction pathways associated with the responses of cells to deformation is essential in the field of tissue engineering, where mechanical environment during growth can affect cell response and the material properties of the tissue construct or living implant.

It is clear that blood rheology plays a major role in the maintenance of the human body and in the development of artificial organs and other medical devices, but our understanding, of that role and of the clinical implications of having an altered blood rheology, is far from being complete.

BIBLIOGRAPHY

Cited References

1. Chien S. Red Cell Deformability and its Relevance to Blood Flow. Ann Rev Physiol 1987;49:177–192.
2. Cokelet GR, et al. The rheology of human human blood measurement near and at zero shear rate. Trans Soc Rheol 1963;7:303–317.
3. Chien S, Usami S, Dellenbeck RJ, Gregersen M. Shear dependent deformation of erythrocytes in rheology of human blood. Am J Physiol 1970; 219:136–142.
4. Sutera SP. The history of Poiseuille's law. Ann Rev Fluid Mech 1993;25:1–19.
5. White GC, Montgomery RR. Clinical aspects of and therapy for von Willebrand disease. In: Hoffman R, et al., editors. Hematology: Basic principles and practice. 3rd ed. New York: Churchill Livingstone Inc.; 2000. 1946–1958.
6. Curry H. Bleeding Disorder Basics. Ped Nursing 2004;30(5): 402–404 and 428–429.
7. Hafezi-Moghadam A, Thomas KL, Cornelssen C. A novel mouse-driven *ex vivo* flow chamber for the study of leukocyte and platelet function. Am J Phys Cell Physiol 2004;286: C876–C892.
8. Sakakibara M, et al. Applicaion of *ex vivo* flow chamber system for assessment of stent thrombosis. Arterioscler Thromb Vasc Biol 2002;22(8):1360–1364.
9. Zimmer R, Steegers A, Paul R, Affeld K, Reul H. Velocities, shear stresses and blood damage potential of the leakage jets of the Medtronic Parallel bileaflet valve. Int J Artif Organs 2000;23(1):41–48.
10. Komorowicz E, Kolev K, Lerant I, Machovich R. Flow rate-modulated dissolution of fibrin with clot-embedded and circulating proteases. Circ Res 1998;82:1102–1108.
11. Blinc A, et al. Flow through clots determine rate and pattern of fibrinolysis. Thromb Haemost 1994;71:230–235.
12. Driessen GK, et al. Effect of reduced red cell "deformability" on flow velocity in capillaries of rat mesentery. Pflügers Arch 1980;388:75–78.
13. Schmalzer EM, Manning RS, Chien S. Filtration of sickle cells: recruitment into a rigid fraction as a function of density and oxygen tension. J Lab Clin Med 1989. 113:727–734.
14. Schmid-Schonbein H, Volger E. Red-cell aggregation and red-cell deformability in diabetes. Diabetes 1976;25(Suppl. 2): 897–902.
15. Schrier SL, Rachmilewitz E, Mohandas N. Cellular and membrane properties of alpha and beta thalassemic erythrocytes are different: implications for differences in clinical manifestations. Blood 1989;74:2194–2202.
16. Reinhart WH, Chien S. Roles of cell geometry and cellular viscosity in red cell passage through narrow pores. Am J Physiol 1985;248(Cell Physiol. 17):C473–C479.
17. Secomb TW, Hsu R. Motion of red blood cells in capillaries with variable cross-sections. J Biomech Eng 1996;118:538–544.
18. Lipowsky HH, Cram LE, Justice W, Eppihimer M. Effect of erythrocyte deformability on *in vivo* red cell transit time and hematocrit and their correlation with *in vitro* filterability. Microvasc Res 1993;46:43–64.
19. Reid HL, Barnes AJ, Lock PJ, Dormandy JA, Dormandy TL. A simple method for measuring erythrocyte deformability. J Clin Pathol 1976;29(9):855–858.
20. Dodds AJ. et al. Haemorheological response to plasma exchange in Raynaud's syndrome. Br Med J 1979. 1186–1187,
21. Boyden S. The chemotactic effect of mixtures of antibody and antigen on polymorphonuclear leucocytes. J Exp Med 1962; 115:453–466.
22. Zigmond SH, Hirsch JG. Leukocyte locomotion and chemotaxis. New methods for evaluation, and demonstration of a cell-derived chemotactic factor. J Exp Med 1973;137:387–410.
23. Bessis M, Mohandas N, A diffractometric method for the measurement of cellular deformability. Blood Cells 1975;1: 307–313.
24. Bull B, Feo C, Bessis M, Behavior of elliptocytes under shear stress in the rheoscope and the ektacytometer. Cytometry 1983;3:300–304.
25. Bareford D, et al. Erythrocyte deformability in peripheral occlusive arterial disease. J Clin Pathol 1985;38:135–139.
26. Erythrocyte deformability in peripheral occlusive arterial disease. J Clin Pathol 1985;38:135–139.
27. Yip R, et al. Red cell membrane stiffness in iron deficiency. Blood 1983;62:99–106.

28. Fung YC. Biomechanics—Circulation 2nd ed. New York: Springer-Verlag; 1997.
29. Pedley TJ. Pulmonary Fluid Dynamics. Ann Rev Fluid Mech 1977;9:229–274.
30. Ferry JD. Viscoelastic Properties of Polymers. 3rd ed. New York: John Wiley; 1980.
31. Whitemore RL. Rheology of the Circulation. New York: Pergamon Press; 1968.
32. Tran-Son-Tay R. Techniques for Studying the Effects of Physical Forces on Mammalian Cells and for Measuring Cell Mechanical Properties. In: Frangos JA, editor. Physical Forces and the Mammalian Cell. New York: Academic Press; 1993;p. 1–59.
33. Tran-Son-Tay R, Beaty BB, Acker DN, Hochmuth RM. Magnetically Driven, Acoustically Tracked Translating Ball Rheometer for Small, Opaque Samples. Rev Sci Instru 1988; 59:1399–1404.
34. Tran-Son-Tay R. A Microrheometer for Studying the Rheological Properties of Sickle Cell Suspensions and Hemoglobin. Proceedings of The Fourth China-Japan-USA-Singapore Conference on Biomechanics. In: Yang G, Hayashi K, Woo SL-Y, Goh JCH, editors. International Academic Publishers; 1995; p. 429–432.
35. Waugh RE, Hochmuth RM. Mechanics and Deformability of Hematocytes. In: Schneck DJ, Bronzino JD, editors. Biomechanics—Principles and Applications. New York: CRC Press; 2202; 227–239.
36. Linderkamp O, Ruef P, Zilow EP, Hoffmann GF. Impaired deformability of erythrocytes and neutrophils in children with newly diagnosed insulin-dependent diabetes mellitus. Diabetologia 1999;42:865–869.
37. Perrault CM, et al. Altered Rheology of Lymphocytes in the Diabetic Mouse. Diabetologia 2004;47:1722–1726.
38. Tsai MA, Frank RS, Waugh RE. Paasive Mechanical Behavior of Human Neutrophils: Effect of Cytochalasin B. Biophys J 1994;66:2166.
39. Thomas SJ, et al. Effects of X-Ray Radiation on the Rheological Properties of Platelets and Lymphocytes. Transfusion 2003;43:502–508.
40. Evans EA,, Yeung A. Apparent viscosity and Cortical Tension of Blood Granulocytes Determined by Micropipet Aspiration. Biophys J 1989;56:151.
41. Hochmuth RM, et al. Viscosity of Passive Neutrophils Undergoing Small Deformations. Biophy J 1993;64:1596–1601.
42. Needham D, Hochmuth RM. Rapid Flow of Passive Neutrophils into a 4 mm Pipet and Measurement of Cytoplasmic Viscosity. J Biomech Eng 1990;112:269.
43. Tran-Son-Tay R, Needham D, Hochmuth RM. Recovery of Passive Neutrophils after Large Deformation: Liquid Drop Model. Proceedings of the ASME, Adv Bioeng 1991;20:421–424.
44. Kan H-C, et al. Effects of Nucleus on Leukocyte Recovery. Ann Biomed Eng 1999;27(5):648–655.
45. Shyy W, et al. Moving Boundaries in Micro-Scale Biofluid Dynamics. Appl Mecha Rev 2001;54:405–453.
46. Schmid-Schoenbein H, et al. A Counter-Rotating "Rheoscope Chamber for the Study of the Microrheology of Blood Cell Aggregation by Microscopic Observation and Microphotometry. Microvasc Res 1973;6:366–376.
47. Tran-Son-Tay R, Sutera SP, Rao PR. Determination of RBC Membrane Viscosity from Rheoscopic Observations of Tank-Treading Motion. Biophy J 1984;46:65–72.
48. Tran-Son-Tay R, Sutera SP, Zahalak GI, Rao PR. Membrane Stress and Internal Pressure in Red Blood Cells Freely Suspended in Shear Flow. Biophys J 1987;51:915–924.
49. Glover S, et al. Phosphorylation of Tyrosine 397 Critically Mediates Gastrin-Releasing Peptide's Morphogenic Properties. J Cellular Physiol 2004;199:77–88.
50. Voldman J, Gray ML, Schmidt MA. Microfabrication in biology and medicine. Annu Rev Biomed Eng 1999;1:401–425.
51. Shyy W, Tran-Son-Tay R, N'Dri N. Micro-Nano Coupling in Biological Systems. In: Harik VM, Luo LS, Salas M, editors. Nano-Scale Mechanics of Solid and Liquid Materials Systems. The Netherlands: Kluwer Academic; 2003.
52. Maddou M. Fundamentals of Microfabrication: The Science of Miniaturization. 2nd ed. Washington (DC): CRC Press; 2001.
53. Xia Y, Whitesides GM. Soft lithography. Ann Rev Mater Sci 1998;28:153–184.
54. Branham ML, et al. Rapid Prototyping of Micropatterned Substrates Using Conventional Laser Printers. J Materials Res 2002;17(7):1559–1562.

Reading List

Adjizian JC, et al. Clinical applications to the Ektacytometer. Clin Hemorheol 1984;4:245–254.
Chien S, et al. Effects of hematocrit and plasma proteins of human blood rheology at low shear rates. J Appl Physiol 1966;21:81–87.

See also CELL COUNTERS, BLOOD; HEMODYNAMICS.

BLOOD, ARTIFICIAL

BRIAN WOODCOCK
University of Michigan
Ann Arbor, Michigan

INTRODUCTION

Blood has such a multitude of physiological functions; it provides a circulating volume to transport substrate and metabolites, and it transports the most valuable of substrates, oxygen. It is an organ intimately involved in the immune system, delivering antibodies and cellular elements to sites of infection. It carries the instruments for coagulation. It is the communication highway for the endocrine system. It is a metabolic organ containing enzyme systems to convert molecules to active and inactive forms. Blood is intimately involved in temperature regulation. The manufacture of an artificial substitute to fulfill all those purposes is beyond the capability of current science. However, several of the functional capabilities of blood have been incorporated into various blood substitutes.

The most basic function of blood is to provide a circulating volume for transportation of substrate and metabolites. Supplementation of intravascular volume with crystalloid and colloid fluids has been a part of medical practice for a century. Recent progress has concentrated on the development of substitute solutions that can transport oxygen. These solutions are known as oxygen therapeutics or red cell substitutes.

Currently, the only available oxygen therapeutic is typed and cross-matched allogeneic human blood. This is made available in the civilian setting by the Red Cross, the American Blood Centers, and the blood banking system. Blood shortages, due to increasing blood usage and declining blood donations (1), are one of the factors driving the

search for alternatives. In the past two decades there has also been an increasing concern regarding the infectious risks of blood borne pathogens (2,3). Awareness of the potential significance of the problem occurred with the increasing risk of human immunodeficiency virus (HIV) transmission from blood transfusion during the 1980s. Improved screening for HIV in the 1990s saw a dramatic improvement in blood safety so that now the risk of contracting HIV from a unit of blood is approaching 1/1,000,000. But there are still substantial concerns regarding not only the risk of contracting acquired immune deficiency syndrome (AIDS) or hepatitis, and also of other infectious diseases newly recognized as possibly being transmissible by transfusion, such as bovine spongiform encephalopathy (mad cow disease) and West Nile virus.

Another concern leading to the need for blood substitutes is that some groups Jehovah's Witnesses, have religious beliefs that cause them to refuse all blood products.

Banked blood is often wasted while active bleeding continues in the surgical setting. While hemorrhage continues, blood products administered are rapidly lost through the site of bleeding. Blood substitutes could be used as a resuscitation bridge until bleeding is controlled (4).

Initial research on artificial blood was lead by the U.S. military, which needed a ready supply of a substitute that could be stored easily and indefinitely in the field and not require typing or cross-matching. These considerations would also make a blood substitute valuable to the emergency medical services in ambulance and helicopter transfers.

All these concerns have stimulated efforts to develop red cell substitutes for use in the routine clinical setting. In an initial approach, prior to World War II, the defense department sought a hemoglobin solution that could be stored indefinitely at room temperature, preferably in a powdered form to be dissolved in normal saline, and that could be transfused without a need for cross-matching. Although the development of a reconstitutable powder has not been feasible, there are several hemoglobin-based products that are in various stages of clinical testing (5).

Other molecules apart from hemoglobin have been assessed for the function of oxygen transportation. Most success has been achieved with emulsions of perfluorochemicals.

Perfluorochemicals are inert liquids, which have a solubility for oxygen and carbon dioxide 20 times that of water. These liquids are immiscible in water and an emulsion form is required to allow them to mix with the recipient's blood after administration. A comparison of the advantages and disadvantages of perfluorocarbon and hemoglobin solutions is shown in Table 1.

Emulsions of perfluorochemicals have completed animal testing and have been investigated in the clinical setting. However, difficulties in their use have been observed, to date they have not been made available for routine use.

Table 1. Comparison of Perfluorocarbon and Hemoglobin Solutions

	Advantages	Disadvantages
Perfluorocarbon Emulsions	High O_2 Solubility Inert Ample supply	Requires high PO_2 Long tissue life Short vascular life Toxicities
Hemoglobin Solutions	Carry O_2 at normal P_aO_2 Unloads like RBCs May be stored dry?	Vasoconstriction Supply Short vascular life Toxicities

CURRENT RISKS OF BANKED BLOOD

Risks of Transfusion

Blood transfusion is safer today than it has ever been, with a death rate for each blood transfusion of ~ 1 in 300,000 (6). Two-thirds of these deaths are due to clerical errors (i.e., the wrong blood given to the wrong patient). Other risks associated with blood transfusion include infection and immune reactions (7,8). However, 20 million blood transfusions are administered each year in the United States, with an impressive safety record (9).

Infection

The risk of transmission of infection includes viral agents, other exotic infectious agents, and the risk of bacterial contamination of blood products. Blood donors with a history of risk factors are excluded from donation. Screening donated units and elimination of units that contain known infectious agents eliminates most of the remaining risk of infection.

HIV

The risk of HIV transmission from blood transfusion has caused the most public concern, though it is difficult to accurately assess the true risk of transmission because it is small and cases can be determined only after a significant period. Initial testing for HIV consisted of antibody testing alone, but this was felt to leave a risk of HIV from individuals who were infected, but had not yet sero-converted. In March 1996, HIV antigen p24 testing was instituted in the United States and only 3 out of 18 million units were identified as being antibody negative and antigen positive over the next 18 months (10). Blood donations are now tested for HIV-1 and HIV-2 (11). The apparent risk of HIV transmission is currently ~ 1 in 1,000,000.

Hepatitis

Hepatitis has a higher prevalence; 1:60,000 for hepatitis B, and 1:103,000 for hepatitis C, but is much less feared by the general public. Antigen screening tests have reduced the risk of post-transfusion hepatitis (B or C) to < 1 in 34,000 (12). Hepatitis G has more recently been recognized, and has a high incidence worldwide of 1 (13)–7% (14). Approximately 2% of blood donors and 15–20% of intravenous drug abusers in the United States have detectable hepatitis G (15). It may be identified by the polymerase chain reaction test, but this has not been implemented as a routine screening test. Fortunately, it appears that the hepatitis G virus is not responsible for non-A, non-B, non-C post-

transfusion hepatitis and the results of infection appear to be minimal (16,17), although there is a weak link between hepatitis G and fulminant hepatitis in rare cases (18).

Other Viruses

Creutzfeldt–Jakob disease (vCJD) can be transmitted by transfer of central nervous system tissue (or extract). However, no cases have been definitively linked to blood transfusion (19).

In the United Kingdom an outbreak of bovine spongiform encephalopathy (BSE) or "Mad Cow Disease" has led to concern that a new variant vCJD could be transferred to the human population through consumption of contaminated beef products. Transmission of BSE by blood transfusion can occur in sheep (20). In the United Kingdom, blood products for transfusion are leucodepleted, which is thought to reduce the risk of transmission of vCJD. The possibility that infection might occur has led the U.S. Food and Drug Administration (FDA) to institute a policy "deferring", that is, declining, blood donations from anyone who has lived in the United Kingdom for a cumulative period of more than 6 months during the years 1980–1996 (21).

West Nile virus transmission has occurred in four patients who received solid organ donations from an infected donor. The organ donor had received blood transfusions from 63 donors, and follow up of those donors showed that one of them was viremic at the time of donation (22).

Bacterial contamination of stored blood is rare (1:500,000), but has a mortality rate of 25–80%. The most common infectious contaminants in red cells are gram-negative species such as *Pseudomonas* or *Yersinia* (23). Platelets are stored at room temperature, allowing rapid bacterial proliferation, and there is a risk of 1:3000–7000 of bacterial infection with these units. Bacterial contamination of platelets is typically with Gram-positive staphylococci.

Immune Reactions

Minor immune reactions, such as febrile reactions, are common and may be discomforting to the patient, but are not associated with significant morbidity, though the transfusion may have to be stopped and the product discarded. The risk of serious immune reactions is small, but present. Most acute hemolytic reactions are due to clerical errors, because cross-matching should predict and prevent these events. However, immune reactions remain the most common cause of fatality associated with transfusions.

Graft versus host disease may occur rarely after transfusion, most commonly after transfusion of nonirradiated blood components to patients with immunodeficiency. Transfusion-associated graft versus host disease has a high mortality and is rapidly fatal. Immunodeficienct patients should receive irradiated units. Immunocompetent individuals may develop graft versus host disease if common histocompatibility leukocyte antigen haplotypes between the donor and recipient prevent destruction of stem cells transfused. This can occur between first-degree family members and therefore, relative-to-patient-directed donations, which are often preferred by patients because of a perceived reduction in risk of infection, may in fact carry an increased risk of initiating transfusion-associated graft versus host disease.

Transfusion-related immunomodulation has been recognized since the mid-1970s, but is not well quantified. Exposure to allogeneic blood can cause both allosensitization and immunosuppression. Studies have demonstrated a beneficial effect of allogeneic blood transfusion on transplant organ survival, but increases in the rates of cancer recurrence and postoperative infection have also been noted (23,24). Leukocyte depletion and removal of plasma may ameliorate the effects of TNF- suppression and interleukin induction (25).

The risk of infection and concerns regarding immune reactions and immunomodulation have been a large incentive to the development of oxygen-carrying colloids.

PERFLUOROCHEMICAL EMULSIONS

Perfluorochemicals (PFCs) are chemically inert liquids with a high solubility for gases. The PFCs that have been used as blood substitutes are 8–10-carbon atom structures that are completely fluorinated. The PFCs are chemically inert, clear, odorless liquids with a density nearly twice that of water. The solubility of PFCs for oxygen is nearly 20 times that of water.

In 1965, Clark and Gollan (26) performed an experiment to see whether an animal could survive if it breathed liquid PFC equilibrated with 1 atm of oxygen. A rat could be submerged beneath this liquid for 30 min and be retrieved in good condition. Respiration of liquid PFC allowed oxygen absorption from the lungs together with CO_2 excretion.

An intravenous injection of PFC is immediately lethal because the injectate is immiscible with water and forms a liquid embolus. An emulsion of an immiscible liquid can, however, mix with water or blood. In 1968, Gehes (27) produced a microemulsion (particle size, 0.1 μm) of a PFC in normal saline. An exchange transfusion could be done, eliminating all normal blood elements, and a rat with a hemoglobin of 0 could survive breathing 100% oxygen.

Because of the inert nature of these compounds, they are not metabolized, but are cleared from the vascular space by the reticulo-endothelial system (RES), and ultimately collected in the liver and spleen. Eventually, the PFC slowly leaves the body as vapor in the respiratory gas.

Perfluorochemical Oxygen Content

Because PFCs transport oxygen by simple solubility, the amount of oxygen they carry is directly proportional to the percentage of PFC in the bloodstream and to the P_aO_2.

Hemoglobin carries most of the oxygen in whole blood and does so in a nonlinear fashion. The plot of oxygen content against P_aO_2 for hemoglobin, known as the oxygen dissociation curve, is seen in Fig. 1. Perfluorochemicals (PFCs) carry oxygen by direct solubility, as does plasma. Because of the shape of the oxygen dissociation curve, hemoglobin is fully saturated and carries little or no more oxygen above a PO_2 of 90 mmHg (11.99 kPa). Because oxygen content dissolved in plasma, or carried by PFCs,

is linearly related to PO_2, additional oxygen is dissolved in the plasma phase or by PFC as oxygen tension increases.

Arterial content of oxygen (C_aO_2) is defined as the volume of oxygen in milliliters (mL) carried by each 100 mL of blood and is defined as follows:

$$C_aO_2 = (Hb \times 1.34 \times SaO_2) + (0.003 \times P_aO_2) \approx 20\,mL/100\,mL \quad (1)$$

(Hb = hemoglobin: S_aO_2, arterial oxygen saturation: P_aO_2, arterial oxygen tension)

With a normal 15 g of hemoglobin and normal P_aO_2 and S_aO_2 values of 90 mmHg (11.99 kpa) and 97%, respectively, an arterial oxygen content of 20 $mL \cdot dL^{-1}$ is obtained.

When blood contains an oxygen carrying PFC, the oxygen content equation requires a third term to represent the contribution from perfluorocarbon.

$$C_aO_2 = (Hb \times 1.34 \times S_aO_2) + (0.003 \times P_aO_2) + (0.057 \times Fct/100 \times P_aO_2) \quad (2)$$

where Fct = fluorocrit, which is the fraction of the blood volume that is PFC (analogous to the Hct).

Note that the solubility factor of PFC in the third term should be 0.06, that is, 20 times that of the solubility factor for oxygen in plasma (0.003), however, it is reduced by the amount of the plasma solubility factor to account for the plasma displaced by the presence of PFC.

The PFCs carry much more oxygen than plasma, but hemoglobin itself is able to carry much more than any PFC. Figure 1 shows that blood with a Hct of 45% will have a C_aO_2 of 20 mL/100 mL at a PO_2 of 100 mmHg (13.33 kPa), but a solution with a Fct of 45% would have an oxygen content of 2.7 mL/100 mL. Because the PFC carries oxygen by direct solubility, it also releases it in direct proportion to the PO_2, unlike the cooperative binding effect of Hg, with which the PO_2 has to fall below the elbow of the curve for oxygen release to occur. The potential contribution of PFCs to oxygen transport can be assessed by looking at the oxygen consumption required by tissues, the Fct and the PO_2 required to allow this quantity of oxygen to be released in the tissues (28).

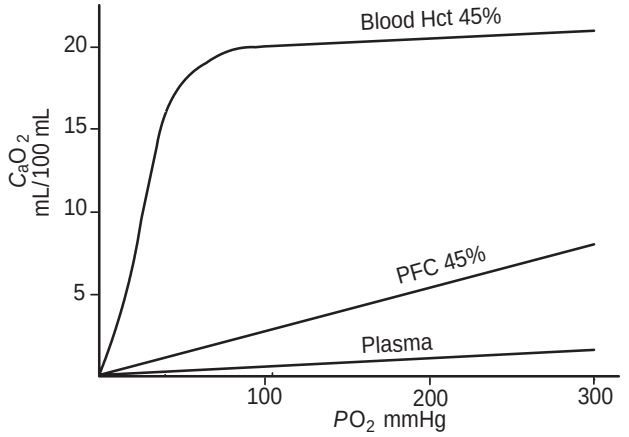


Figure 1. Oxygen content plotted against PO_2 for whole blood, plasma, and a perfluorochemical emulsion, with a 45% content of PFC. Hemoglobin saturates at a PO_2 of 100 mmHg (13.33 kPa), but the curve continues to rise because of the dissolved oxygen in plasma, so the whole blood and plasma lines are parallel above a PO_2 of 100. The line for perflubron is similar to plasma, but has a higher slope because of the greater affinity for oxygen (Hct = hemocrit).

Mixed venous blood has an oxygen content of 15 mL/100 mL; therefore 5 mL/100 mL of oxygen is consumed in the periphery.

If we assume a mixed venous oxygen tension (P_vO_2) level of 40 mmHg (5.33 kPa) and an oxygen extraction of 5 mL/dL, a bloodless animal could survive with a Fct of 45% with a P_aO_2 of 235 mmHg (31.33 kPa). Any increase in P_aO_2 above this value would raise the P_vO_2 by the same amount (Fig. 2). The elevated P_vO_2 in these circumstances could have the beneficial effect of increasing the pressure gradient for oxygen diffusion from the vascular space into the tissues and cells, theoretically increasing tissue oxygenation.

It is, however, difficult to manufacture a 45% emulsion of PFC. In the late 1970s, the Green Cross Corporation in Japan developed a product called Fluosol DA 20%. This solution contains only 10% PFC (Fluosol DA 20% is 20% by weight, 10% by volume). To supply 5 mL/100 mL of oxygen consumption and a P_vO_2 of 40 mmHg (5.33 kPa), a bloodless animal with a fluocrit of 10% would require a P_aO_2 of 920 mmHg (122.65 kPa).

In practice, this would make it difficult to completely replace the blood with PFC emulsion. Although the emulsion is cleared from the vascular space within 24 h, the long tissue half-life of a PFC in the body (months to years), make it unfeasible to continuously redose the patient.

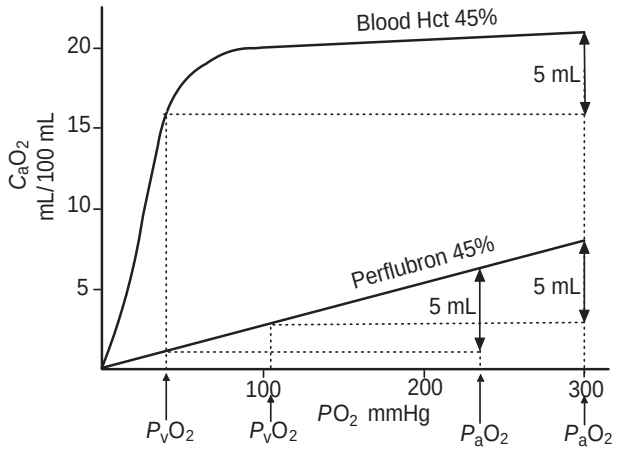


Figure 2. At an arterial PO_2 (P_aO_2) of 300 mmHg (13.33 kPa) blood has an oxygen content of 21 mL/100 mL. If 5 mL/100 mL are extracted the venous PO_2 (P_vO_2) will be < 50 mmHg. A PFC with a Fct of 45% could carry enough oxygen at a P_aO_2 of 235 to deliver 5 mL and give a similar P_vO_2. Increasing the P_aO_2 to 300 mmHg (39.99 kPa), increases the oxygen carried by the PFC so the P_vO_2 will be > 100 mmHg (13.33 kPa) after delivery of 5 mL O_2. (Adapted, with permission, from Woodcock BJ, Tremper KK. Red Blood Cell Substitutes. In: Evers AS, Maze M, editors. Anesthetic Pharmacology, Physiological Principles and Clinical Practice: A Companion to Miller's Anesthesia. Philidelphia: Churchill Livingstone; 2004.)

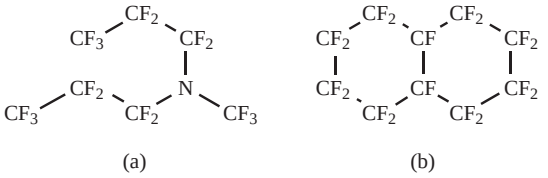


Figure 3. Structures of the constituents of Fluosol DA; perfluorodecalin (a) and perfluortripropylamine (b).

CLINICAL STUDIES WITH PFCS

Fluosol DA

Fluosol DA 20% was developed in the 1970s and was an emulsion composed of two perfluorochemicals; perfluorodecalin, which has emulsion stability and perfluorotripropylamin, which has a shorter half-life (29) (Fig. 3). The surfactant used to maintain this emulsion of Fluosol DA was Pluronic F68, an emulsifier used in other medical products. The product needed to be frozen to maintain stability until used.

Initial clinical studies were in patients who were actively bleeding, required surgery, and refused blood transfusion. A 20-$mL \cdot kg^{-1}$ body weight of Fluosol DA 20% PFC emulsion was transfused preoperatively with a maximum of one additional dose postoperatively (29,30). The patients achieved a maximal fluorocrit of $< 3\%$ with an intravascular half-life of ~ 19 h. Subsequent studies confirmed the oxygen-carrying contribution of the PFC, but could not show a beneficial effect on patient outcome (31).

Fluosol DA 20% gained FDA approval, not for use as a red cell substitute but for intracoronary infusion in patients undergoing angioplasty. Fluosol is no longer being manufactured because of low demand.

Perflubron

Second generation PFC emulsions have been developed. One of these, Oxygent, employs perfluoro-octylbromide or perflubron (32) (Fig. 4). This PFC has one bromine replacing fluorine, making it radiopaque. The emulsion (Oxygent: Alliance Pharmaceutical, San Diego, CA) contains 60% PFC by weight, or 30% by volume. It is emulsified with lecithin (egg yolk phospholipids) and is stable at room temperature for > 6 years.

Because of a short intravascular half-life and the need for a high F_iO_2, it has been suggested that these PFC solutions should be used in conjunction with acute normovolemic hemodilution. In this setting, patients who are expected to have significant surgical blood loss should have two to four units of blood removed immediately before surgery. Initial volume expansion is with crystalloid or colloid. As surgical bleeding continues, the PFC is administered at the first transfusion trigger, allowing for adequate oxygen delivery at low hemoglobin levels. When surgical bleeding stops, the patient would receive the previously harvested autologous blood.

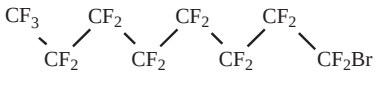


Figure 4. Structure of perfluoro-octylbromide (Perflubron, Oxygent).

The European Perflubron Emulsion Study Group looked at this technique of normovolemic hemodilution in orthopedic surgery. Perflubron was administered in response to transfusion triggers of tachycardia, hypotension, increased cardiac, and decreased mixed venous PO_2. The triggers for transfusion were at least as effectively reversed by perflubron, with 100% oxygen ventilation, as by autologous transfusion or colloid administration (33). Patients who received perflubron had higher mixed venous PO_2 levels, which continued for longer than the other treatment groups. Allogeneic blood transfusion requirements were not reduced by perflubron administration.

A phase III cardiac surgery study of perflubron was voluntarily suspended in 2001 because of an increased incidence of stroke, but it has not been determined if this was due to PFC administration.

A phase III study of perflubron in noncardiac surgery, again from the European Perflubron Emulsion Study Group (34), in 492 patients showed an intraoperative reduction in blood transfusion requirement, but there was no significant reduction by time of the postoperative period. Of the patients in the intent-to-treat population, 33% experienced low (< 20 $mL \cdot kg^{-1}$) blood loss. When these patients were excluded to leave a protocol defined target population, the patients treated by acute normovolemic hemodilution (ANH) and PFC administration experienced a reduction in transfusion that remained significant throughout the study (3.4 vs. 4.9 units). The PFC group also had a significantly greater avoidance of transfusion (26 vs. 16%). The clinical relevance of these findings is hard to elucidate. The control group did not undergo ANH and had higher transfusion triggers during the operative period, it is difficult to tell if the benefit accrued was due to the PFC or the hemodilution technique (35).

Microcirculation and Oxygen-Carrying Colloids

Oxygen-carrying colloids differ from red cells in terms of size. Free hemoglobin molecules in solution range in size from 68,000 to 500,000 Da, and the fluorochemical emulsion particles are $\sim 0.2\,\mu m$. These sizes are within the range of many of the plasma proteins, and like plasma proteins, PFCs are able to cross the capillary basement membrane and participate in extravascular circulation. Animal data have shown that these oxygen-carrying colloids leave the intravascular space and rejoin it through lymphatic circulation (36). Thus, these colloids may be able to provide increased oxygen delivery to the extravascular space (37,38). In addition, studies on microcirculation have shown that plasma flow continues through capillaries even when the capillary is closed to red blood cells. Oxygen delivery to tissues can be provided, even though red cells are not present in the capillaries, through the circulation of oxygen-carrying colloids.

This potential application of blood substitutes, as "therapeutic oxygenating agents" for ischemic tissue, has been

investigated for myocardial ischemia (39–41) and for enhancing the effectiveness of radiation therapy and chemotherapy of ischemic tumors by rendering the tumor hyperoxic (42,43).

Perfluourochemicals have been used in other circumstances as oxygen-carrying molecules. The PFCs have been used for cardiplegia (44) and preservation of transplanted organs (45–47). The PFCs have also been studied in models of myocardial and cerebral infarcts to minimize infarct size (45,48,49).

Liquid Ventilation with PFC

Perfluorochemicals have been used for liquid ventilation in the treatment of acute respiratory distress syndrome (ARDS). The PFCs bind oxygen and carbon dioxide avidly. Perflubron (LiquiVent, Alliance Pharmaceutical Corp., San Diego, CA) can be instilled into the endotracheal tube of a ventilated patient with ARDS until a fluid level is seen outside the patient. The ET is then connected to the normal ICU ventilator and sufficient gas transfer occurs across the PFC–gas interface to allow oxygenation of the patient and CO_2 clearance (50,51). This PFC has excellent surfactant properties and is possibly able to stent open alveoli (leading it to be termed "liquid PEEP" or "PEEP in a bottle") (52). There are also benefits of increased secretion clearance and possible antiinflammatory effects (53,54). Studies to date have not shown any benefit over conventional ventilation (55).

Future of PFCs

The PFCs have inherent limitations of a short endovascular half-life and the requirement for high inspired oxygen. These problems limit the use of PFC emulsions to acute settings in which supplemental oxygen is readily available. It is yet to be seen whether PFCs can have a useful role as blood substitutes.

HEMOGLOBIN SOLUTIONS

Hemoglobin Solutions: Oxygen Content

When a solution of free hemoglobin (FHb) has the same $P50$ as blood (27 mmHg 3.59 kPa), there is no difference between the hemoglobin solution oxygen-content curve and the curve for normal whole blood.

$$C_aO_2 = (\mathrm{FHb} \times 1.34 \times \mathrm{F}S_aO_2) + (\mathrm{Hb} \times 1.34 \times S_aO_2) + (0.003 \times P_aO_2) \quad (3)$$

[FHb is the concentration of free hemoglobin solution in blood ($g \cdot dL^{-1}$), and FS_aO_2, is the saturation of FHb.]

In clinical practice, arterial content of oxygen can be calculated using a single term for hemoglobin and saturation. A spectrophotometrically measured total hemoglobin should be used, which will measure total hemoglobin present, whereas the hematocrit only measures red blood cell hemoglobin. The saturation measured by an oximeter gives a mean saturation of both forms of hemoglobin because the device measures the amount of oxyhemoglobin and divides it by total hemoglobin to achieve the calculated saturation.

Formulation of Hemoglobin Solutions

A solution of hemoglobin from lyzed human red cells is unusable as a blood substitute for a variety of reasons. Hemoglobin outside the red cell membrane loses its tetrameric form and breaks down into dimers. The abundance of dimers in plasma has a pronounced oncotic effect creating an excessively high colloid oncotic pressure (41). This would draw fluid from the extracellular space and would cause an increase in circulating blood volume. The dimers have a molecular weight of 32,000 Da and are able to cross the renal glomerular basement membrane leading to a potent osmotic diuretic effect. The loss of oxygenated hemoglobin in the urine gave the early solutions the reputation of being "red mannitol".

The loss of 2,3-DPG, which is normally maintained inside the RBC, reduces the P50 of hemoglobin to 12–14 mmHg (1.59–1.86 kPa) from a normal level of 26. This shifts the oxygen dissociation curve markedly to the left, meaning that the free hemoglobin will avidly bind oxygen during passage through the lungs but will not release it in the peripheral tissues unless the PO_2 is extremely low.

The early attempts at making a hemoglobin solution used resuspended hemoglobin filtered from lyzed, outdated, human blood. This solution caused a high incidence of renal failure, which proved not to be due to the free hemoglobin, but due to the "stroma" of residual red blood cell elements left after cell lysis (56).

Several types of hemoglobin solution have been developed, and have taken different approaches to these problems. One product is produced from modified polymerized bovine hemoglobin (Hemopure, HBOC-201, Biopure, Cambridge, USA) (57,58). The dimer form is polymerized to form a larger roughly octomeric molecules (Fig. 5). It is

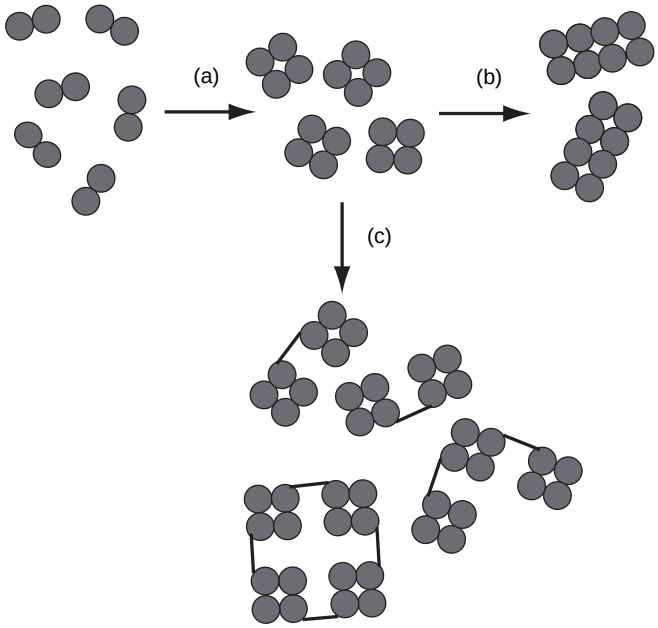


Figure 5. Hemoglobin dimer subunits can be polymerized (a) to form tetramers. These can be further polymerized (b) to form octamers or larger units, or can be cross-linked (c) to increase molecular size.

Table 2. Hemoglobin Solutions in Clinical Trials

Product (Manufacturer)	Configuration	*P*50	Status
	Bovine Hemoglobins		
HBOC201 Hemopure (Biopure)	Polymerized with glutaraldehyde	34 mmHg	Licensed in South Africa and for veterinary use. FDA review; further studies required
PEG-hemoglobin (Enzon)	PEG (polyethylene glycol) conjugated noncross-linked, encapsulated	20 mmHg	Phase 1 melanoma studies completed Now discontinued
	HumanHemoglobins		
Polyheme (Northfield)	Polymerized, tetramer-free	30 mmHg	Phase 3 study, before FDA review
Hemolink (Hemosol)	Polymerized with *o*-raffinose	32 mmHg	Under review United Kingdom and Canada
DCLHb (Baxter)	Cross-linked with diaspirin	32 mmHg	Discontinued
PHP (Curacyte)	Pyridoxylated		Studied in SIRS
	Recombinant Hemoglobins		
Optro (Somatogen)	Genetically fused	17 mmHg	Discontinued

then filtered to remove the smaller molecular weight tetrameric hemoglobin molecules. This reduces the oncotic pressure and prevents passage of the molecule though the glomerulus into the urine. Polymerization also increases the *P*50 into the normal range and allows the hemoglobin solution to release oxygen in the tissues (41).

Other products have been developed from outdated human blood: Cross-linking hemoglobin molecules with pyridoxal-5-phosphate (Fig. 5), which acts as an artificial 2, 3-DPG, can increase the *P*50. This technique is used in PolyHeme polymerized hemoglobin solution (Northfield Pharmaceutical, Chicago, IL) (59–61), and pyridoxylated hemoglobin polyoxyethylene conjugate or PHP hemoglobin (62). Diaspirin Cross-Linked Hemoglobin, DCLHb (Baxter Healthcare, Chicago, IL) (63,64) and Hemolink (Hemosol, Toronto, Ontario, Canada) also have a *P*50 within the normal range and have an increased size due to cross-linking or polymerization.

The size of the hemoglobin molecule can also be increased by attaching the dimer to a large non-hemoglobin molecule, for example, polyethylene glycol used in PEG–hemoglobin (Enzon, Piscataway, NJ). A final method of increasing molecular size is to encapsulate hemoglobin in phospholipid vesicles or liposomes (65).

Hemoglobin cannot only be obtained from outdated human blood, but also from bovine blood. There is a tremendous supply of bovine blood, since nearly 1 million units/day are produced as a byproduct of meat production. Bovine hemoglobin does not normally require 2, 3-DPG, and maintains a *P*50 in the range of 32 mmHg (4.26 kPa), even as a dimer (66). Concern over the spread of BSE or variant vCJD may impact the development of bovine products (67). The FDA has restricted the import of bovine products from Europe because of the outbreak of Mad Cow disease there (68).

Recombinant hemoglobin has been used as an alternative source of hemoglobin. Optro (Somatogen), was engineered to have a *P*50 in the normal range. However, manufacture has been discontinued (67).

The hemoglobin solutions that have undergone clinical trial are listed in Table 2. Since these solutions are produced from different hemoglobin sources, and have different sizes and different *P*50 values, each needs to be evaluated as a separate drug with its own effectiveness and toxicity profile.

All of these products have a relatively short intravascular half-life compared to hemoglobin contained in red blood cells, and are cleared from the vascular space by the reticuloendothelial system with a half-life of $\sim$24 h.

Hemodynamic Effects of Hemoglobin Solutions

Pulmonary and systemic hypertension have been observed in studies in animals (63,69–72) and human clinical studies (36,38,57,64,73–78). The cause of this appears to be related to the nitric oxide (NO) scavenging properties of hemoglobin. Endothelium derived relaxant factor (EDRF) was identified as NO in the 1990s and its role in controlling vascular resistance was elucidated. Nitric oxide is produced in the endothelial cells of blood vessel walls, and produces smooth muscle relaxation, thereby causing vasodilatation. As blood flow increases, nitric oxide is carried away, reducing its concentration and causing vasoconstriction (79). Binding of nitric oxide to hemoglobin plays an important role in its removal, and thereby the control of vascular tone. Free hemoglobin binds nitric oxide more avidly than hemoglobin within the red cell and nitric oxide clearance is increased (80). This leads to hypertension in the pulmonary and systemic vascular beds. Smaller hemo-

globin molecule size appears to increase NO clearance, and pulmonary hypertension becomes more problematic. The severity of this side effect varies with the different forms and preparations of hemoglobin.

The observed vasoconstriction with hemoglobin solutions led to the evaluation of DCLHb solution as an "all-in-one" therapeutic strategy in vasodilated shock states (38) as a vasopressor in critically ill patients with septic shock. In septic shock or systemic inflammatory response syndrome, the vasodilated state may be due to excessive synthesis of NO. The scavenging effect of the hemoglobin solution may be beneficial in restoring vascular tone in this setting (37). Tissue perfusion may be improved because of the small size of the hemoglobin molecules compared to red cells, improving oxygen delivery through the microcirculation. In a mouse model of gram-negative sepsis, hemoglobin solution was associated with increased circulating tumor necrosis factor and increased lethality (81). In pigs with septic shock, DCLHb administration restored blood pressure and allowed a reduction in dopamine infusion being used for resuscitation (82). Further work is required to determine the utility of hemoglobin solutions in critical illness (83).

Although NO scavenging may play a major role in the vasoconstriction seen with hemoglobin solutions it may not explain the whole picture. Hemoglobin molecules with similar rates of NO binding may have strikingly different degrees of vasoconstriction. The fall in PO_2 in terminal arterioles may play a large role in the autoregulation of microvascular circulation by causing vasoldilation. Hemoglobin solutions may deliver excess oxygen to the terminal capillaries causing vasoconstriction and paradoxically reducing oxygen delivery to the tissue (84).

This proposed effect on microcirculation has led to the adoption of a "counterintuitive" approach to hemoglobin solution development. An anemic but hyperoncotic solution with a very low $P50$ could delay oxygen release and prevent vasoconstriction (85). Such a solution has been developed in MalPEG-Hb (maleimide-activated polyethylene glycol conjugated hemoglobin). This MalPEG-Hb solution has a $P50$ of 5.5 mmHg and has been shown to improve microvascular circulation in hypovolemic shock in hamsters (86).

Duration of Action of Hemoglobin Solutions

Just as free hemoglobin is scavenged following intravascular hemolysis, the RE system scavenges hemoglobin solution from the blood stream. The effective intravascular life of the hemoglobin solution is $\sim$ 12–48 h, depending on the dose (66). Therefore, hemoglobin solutions, like the PFCs, will not have a role in maintaining oxygen-carrying capacity in chronically anemic patients. They may be used in acute, limited blood-loss situations as resuscitative fluids, especially in combination with perioperative autologous hemodilution to minimize loss of the patients own blood (87,88).

Hematopoietic Effect of Hemoglobin Solutions

Hemoglobin solutions may have a profound effect in stimulating erythropoiesis. Studies in which animals are hemodiluted to a low Hct, and given a transfusion of hemoglobin solution, demonstrate a pronounced level of red blood cell production exceeding that expected even with the administration of exogenous erythropoietin (36).

Free iron liberated when hemoglobin solutions are broken may stimulate erythropoiesis, reducing the requirement for subsequent red cell transfusion (89,90). Transfusion of diaspirin cross-linked hemoglobin (DCLHb) to transfuse cardiac surgery patients in the postbypass period resulted in 19% of patients not requiring packed red blood cell transfusion, although the DCLHb was rapidly cleared from the circulation (77). After acute normovolemic hemodilution (ANH) with HBOC-201 in human volunteers there were increases in serum iron, ferritin, and erythropoietin that did not occur following ANH with Ringer's lactate solution (91).

Other Uses of Hemoglobin Solutions

Hemoglobin solutions have an oxygen delivery curve similar to that of hemoglobin in red blood cells. At low temperatures the curve is shifted to the left, preventing release of the bound oxygen, they therefore may not be of any benefit as a constituent of cardioplegia solutions or organ preservatives used for transplanted organs.

Artifical oxygen carriers have a potential for abuse by elite athletes, and international sporting federations have already added the class of agent to their banned substance lists (92).

Clinical Trials: Hemoglobin Solutions

HBOC-201 (Hemopure). The bovine product HBOC-201 (Hemopure, Biopure, Cambridge, MA) is produced from modified polymerized bovine hemoglobin and has been studied in patients undergoing abdominal aortic aneurysm surgery. Treatment with the hemoglobin solution produced an increase in pulmonary and systemic vascular resistance and an associated decrease in cardiac output (57). In another study in aortic surgery HBOC-201, 27% of patients did not need blood transfusion, but the median transfusion requirement was not decreased. Mean arterial blood pressure increased by 15% in this study (75). Vasoconstriction and hypertension have been noted in many studies with HBOC-201. A study looking at preoperative hemodilution with bovine HBOC-201 before liver resection showed an increase in systemic vascular resistance and a fall in cardiac output (73). More recently, a study by Wahr et al. (36) found this product to be useful in reducing the amount of allogeneic blood in operative patients. Increases in pulmonary or systemic resistance were not seen, but a mild increase in blood pressure occurred.

In postoperative cardiac surgery patients, administration of up to 1000 mL HBOC-201 prevented packed red blood cell administration in 34% of patients who would otherwise have received it (93). Although the HBOC-201 had a short duration in the circulation, Hct was restored rapidly, perhaps due to a hematinic effect. The HBOC-201 patients had a slightly greater increase in blood pressure after transfusion.

HBOC-201 has been administered to patients with sickle cell anemia and may have a role in the treatment of vasoocclusive or aplastic crises in this disease (94,95).

The HBOC-201 can be used instead of PRBCs to transfuse patients with severe autoimmune hemolytic anemia, the hemoglobin solution does not have the surface antigens associated with red blood cells (96). In the presence of a Hct of 4.4% the hemoglobin solution reversed both lactic acidosis and myocardial ischemia.

In South Africa, HBOC-201 (Hemopure) has been licensed to treat anemia in surgical patients; this is the first time a hemoglobin solution has reached the market in humans. It is also licensed, as Oxyglobin, for veterinary use in the United States and Europe. The FDA approval for human use in the United States has been delayed pending a request for further information and animal testing.

Diaspirin Cross-Linked Hemoglobin, DCLHb

The DCLHb (Baxter Healthcare, Chicago, IL) has been developed from outdated human blood. Clinical studies to date have had varying results with DCLHb.

The DCLHb effectively scavenges NO and has been noted to increase pulmonary and systemic vascular resistance in the hemodilution model in animals to the point of reducing cardiac index and oxygen delivery relative controls (63). Pulmonary and systemic hypertension with rises in SVR and PVR also occurred in human studies of DCLHb given after cardiac surgery, and this led to decreased cardiac output compared to patients given red cell transfusion (77). Another study (76) showed that DCLHb could be used to reduce the number of patients requiring perioperative packed red blood cell (PRBC) transfusions following vascular, orthopedic, and abdominal surgery compared with patients randomized to PRBC transfusions. However, the total PRBC and other blood product requirements of the two groups were similar over the subsequent week. Side effects of hypertension, jaundice and hemoglobinuria were noted. One death from respiratory distress syndrome was attributed to DCLHb, and the study was terminated early.

This solution was also investigated in a randomized study of patients with acute ischemic stroke (64). Patients receiving the hemoglobin solution had more deaths, serious adverse outcomes, and worse outcome scale scores.

Studies in patients with hemorrhagic shock in the United States and Europe were discontinued because of increased mortality in the U.S. study, and a lack of benefit with DCLHb administration in the European study (97).

Because of these results Baxter has discontinued further work with DCLHb. He subsequently acquired Somatogen, manufacturer of Optro, a first generation recombinant hemoglobin, development of which has since been discontinued.

PolyHeme

PolyHeme (Northfield Pharmaceutical, Chicago, IL) has been used in trials treating acute trauma patients. Gould et al. administered 1–20 units of PolyHeme to 171 patients with urgent blood loss; 81 patients received 5 or more units and 34 received 10–20 units (98). Overall there was a mortality of 10.5 compared to 16% of historical controls who declined blood transfusion because of religious reasons during surgery, and had Hb levels < 8 g/dL. However, there was no comparison with controls receiving transfusion of allogeneic packed red blood cells. An earlier study, also by Gould, randomly assigned 44 trauma patients to either receive blood (23 patients) or up to 6 units of PolyHeme (21 patients) (60). There were no adverse effects on pulmonary and systemic vascular resistance or cardiac output from the administration of PolyHeme. There were no differences in patient outcomes, although there was a reduced need for red blood cells in the PolyHeme group at day 1, which was no longer seen by day 3. The lack of effect of PolyHeme on systemic and peripheral vascular resistance is attributed to its manufacturing process, which filters out the smaller tetrameric hemoglobin (59,60). It is speculated that the smaller size hemoglobin elements can defuse through the vessel wall, increasing NO scavenging and producing vasoconstriction.

Studies have found an intravascular half-life of 24 h of PolyHeme, which is longer than that found in previous studies that found a halflife in the range of 9–12 h. It may be that the half-life of these products is dose dependent; studies showing a larger dose produces a longer half-life.

PolyHeme is currently being assessed in a large multicenter study compared to saline for trauma patients. An application to the FDA for approval may follow this study and PolyHeme could have a role in future clinical practice.

Hemolink

Hemolink (Hemosol, Toronto, Ontario, Canada) is an O-raffinose cross-linked human hemoglobin, which has been shown to cause vasoconstriction in animals (70). As with other cross-linked hemoglobin solutions, the hemodynamic effects are less pronounced than with unmodified hemoglobin solutions (99) but exceed that of polymerized hemoglobin solutions.

A Phase I study in healthy human volunteers showed the solution was well tolerated apart from some moderate to severe abdominal pain, which occurred in all subjects at higher doses. Blood pressure rose by 14% following administration, and with higher doses this elevation lasted 24 h (78). The findings of a Phase II study in coronary artery bypass (CAB) surgery (74) reported a 7–10% increase in blood pressure, which was not statistically significant, and a reduction in the number of patients requiring red cell transfusion from 57 to 10%. A further report by the same group using intraoperative autologous donation and volume replacement with Hemolink or pentastarch in CAB surgery abolished the need for intraoperative transfusion (0 vs. 17% in the pentastarch group) (100). The reduction in transfusion requirement continued at 1 day (7 vs. 37%) and 5 days (10 vs. 47 %) after surgery. Adverse effects included hypertension (43 vs. 17%) and atrial fibrillation (37 vs. 17%).

A similar study in CAB patients using intraoperative autologous blood donation (IAD) and volume replacement with Hemolink or pentastarch showed a reduction in transfusion from 76% in the pentastarch group to 56% with Hemolink (101). This was compared to an historical group of patients in whom IAD was not used, who required transfusion in 95% of operations. Hypertension was again noted in the Hemolink treated group.

Hemolink is under review for approval by the drug agencies of the United States, Canada, and the United Kingdom.

PEG Hemoglobin

The PEG Hemoglobin (Enzon, Piscataway, NJ) is a bovine hemoglobin conjugated with polyethylene glycol. This increases the molecular size without using cross-linking between molecules. Retention in the circulation is increased as the conjugated molecule does not cross the glomerular basement membrane. Administration in dogs resulted in no elevation of blood pressure and it was well tolerated (102).

The PEG hemoglobin has been used to sensitize tumors to chemotherapy (103) and radiotherapy in rodents (104). The small molecular size, compared to red cells, improves microvascular oxygenation, and tumors that are hypoxic become more responsive to chemotherapy and radiotherapy.

The PEG hemoglobin has also been used in rabbits, in the preservative perfusate, and for protection of transplanted hearts during the ischemic period with improvement in cardiac function post-transplant (105).

PHP

Pyridoxilated hemoglobin polyoxyethylene conjugate or PHP (Curacyte, Chapel Hill, NC) is a human hemoglobin solution, cross-linked by pyridoxal-5-phosphate, which is being developed as a NO scavenger for use in septic shock and systemic inflammatory response syndrome (62). A Phase II study has been completed and a Phase III study is in progress looking at PHP for the treatment of NO induced shock.

Encapsulated Hemoglobins

The problems of small hemoglobin molecule size, leading to NO scavenging, high oncotic pressures, and osmotic diuresis can be countered by encapsulating the hemoglobin. This mimics the natural presentation of hemoglobin in whole blood and is sometimes referred to as "neo red cells" (65). Most work has used liposome encapsulated hemoglobin (LEH), but biodegradable polymer microcapsules have also been used (106). Circulation time of the hemoglobin is increased by encapsulation and can be increased, from 18 to 65 h, by polyethylene glycol (PEG) modification (107), these long-lasting derivatives have been named "stealth" liposomes. However, there is significant accumulation of liposomes in the liver and spleen, when LEH is given, causing vacuolization seen on liver biopsy. Liver transaminases may also be elevated (41).

BIBLIOGRAPHY

Cited References

1. Epstein JS. The US blood supply. Am Fam Phys 2000;61: 549–550.
2. American Society of Anesthesiologists Task Force on Blood Component Therapy, Practice Guidelines for blood component therapy. Anesthesiology 1996;84:732–747.
3. Spahn DR, Casutt M. Eliminating blood transfusions: new aspects and perspectives. Anesthesiology 2000;93:242–255.
4. Cohn SM. Blood substitutes in surgery. Surgery 2000;127: 599–602.
5. Ketcham EM, Cairns CB. Hemoglobin-based oxygen carriers: development and clinical potential. Ann Emerg Med 1999; 33:326–337.
6. Myhre BA, Bove JR, Schmidt PJ. Wrong blood—a needless cause of surgical deaths. Anest Analg 1981;60:777–778.
7. Nichollis MD. Transfusions: morbidity and mortality. Anaesth Intensive Care 1993; 15–19.
8. Sazama K. Reports of 355 transfusion-associated deaths: 1976 through 1985. Transfusion (Paris) 1990;30:583–590.
9. Newman RJ, Podolsky D, Loeb P. Bad blood. US News World Rep 1994;116:68–70.
10. Lackritz EM. Prevention of hiv transmission by blood transfusion in the developing world: achievements and continuing challenges. AIDS 1998;12:81–86.
11. Chamberland M, Khabbaz RF. Emerging issues in blood safety. Inf Dis Clin North Am 1998;12:217–229.
12. Dodd RY. The risk of transfusion-transmitted infection. N Engl J Med 1992;327:419–421.
13. Yoshikawa A, Fukuda S, Itoh K, Kosaki N, Suzuki T, Hirakawa K, Nakao H, Inoue T, Fukuda M, Okamoto H. Infection with hepatitis g virus and its strain variant, the gb agent (gbv-c), among blood donors in japan. Transfusion (Paris) 1997;37:657–663.
14. Tacke M, Kiyosawa K, Stark K, Schlueter V, Ofenloch-Haehnle B, Hess G, Engel AM. Detection of antibodies to a putative hepatitis g virus envelope protein. Lancet 1997;349:318–320.
15. Fiebig EW, Busch MP. Emerging infections in transfusion medicine. Clin Lab Med 2004;24:797.
16. Alter MJ, Gallagher M, Morris TT, Moyer LA, Meeks EL, Krawczynski K, Kim JP, Margolis HS. Acute non-a-e hepatitis in the united states and the role of hepatitis g virus infection. Sentinel counties viral hepatitis study team. N Engl J Med 1997;336:741–746.
17. Alter HJ, Nakatsuji Y, Melpolder J, Wages J, Wesley R, Shih JW, Kim JP. The incidence of transfusion-associated hepatitis g virus infection and its relation to liver disease. N Engl J Med 1997;336:747–754.
18. Karayiannis P, Thomas HC. Current status of hepatitis g virus (gbv-c) in transfusion: is it relevant? Vox Sang 1997;73:63–69.
19. Ricketts MN, Cashman NR, Stratton EE, ElSaadany S. Is creutzfeldt-jakob disease transmitted in blood? Emerg Infect Dis 1997;3:155–163.
20. Houston F, Foster JD, Chong A, Hunter N, Bostock CJ. Transmission of bse by blood transfusion in sheep. Lancet 2000;356:999–1000.
21. Mitka M. Blood groups differ on donor deferral. JAMA 2001;285:1694–1695.
22. Iwamoto M, Jernigan DB, Guasch A, Trepka MJ, Blackmore CG, Hellinger WC, Pham SM, Zaki S, Lanciotti RS, Lance-Parker SE, DiazGranados CA, Winquist AG, Perlino CA, Wiersma S, Hillyer KL, Goodman JL, Marfin AA, Chamberland ME, Petersen LR. West Nile Virus in Transplant Recipients Investigation Team, Transmission of west nile virus from an organ donor to four transplant recipients. N Engl J Med 2003;348:2196–203.
23. Blumberg N, Triulzi DJ, Heal JM. Transfusion-induced immunomodulation and its clinical consequences. Transfus Med Rev 1990;4:24–35.
24. Bordin JO, Blajchman MA. Immunosuppressive effects of allogeneic blood transfusions: implications for the patient with a malignancy. Hematol Oncol Clin North Am 1995;9: 205–218.
25. Biedler AE, Schneider SO, Seyfert U, Rensing H, Grenner S, Girndt M, Bauer I, Bauer M. Impact of alloantigens and

storage-associated factors on stimulated cytokine response in an in vitro model of blood transfusion. Anesthesiology 2002;97:1102–1109.
26. Clark Jr LC, Gollan F. Survival of mammals breathing organic liquids equilibrated with oxygen at atmospheric pressure. Science 1966;152:1755–1766.
27. Gehes RP, Monroe RG, Taylor K. Survival of rats having red cells totally replaced with emulsified fluorocarbon. Fed Pro 1968;27:384.
28. Woodcock BJ, Tremper KK. Red Blood Cell Substitutes. In: Evers AS, Maze M, editors. Anesthetic Pharmacology: Physiological Principles and Clinical Practice: A Companion to Miller's Anesthesia. Philidelphia: Churchill Livingstone; 2004.
29. Tremper KK, Friedman AE, Levine EM, Lapin R, Camarillo D. The preoperative treatment of severely anemic patients with a perfluorochemical oxygen-transport fluid, Fluosol-DA. N Engl J Med 1982;307:277–283.
30. Tremper KK, Levine EM, Waxman K. Clinical experience with Fluosol-DA (20%) in the United States. Int Anesthesiol Clinic 1985;23:185–197.
31. Gould SA, Rosen AL, Sehgal LR, Sehgal HL, Langdale LA, Krause LM, Rice CL, Chamberlin WH, Moss GS. Fluosol-DA as a red-cell substitute in acute anemia. N Engl J Med 1986;314:1653–1656.
32. Keipert PE, Faithfull NS, Bradley JD, Hazard DY, Hogan J, Levisetti MS, Peters RM. Oxygen delivery augmentation by low-dose perfluorochemical emulsion during profound normovolemic hemodilution. Adv Exp Med Biol 1994;345: 197–204.
33. Spahn DR, van Brempt R, Theilmeier G, Reibold JP, Welte M, Heinzerling H, Birck KM, Keipert PE, Messmer K, Heinzerling H, Birck KM, Keipert PE, Messmer K. Perflubron emulsion delays blood transfusions in orthopedic surgery. European perflubron emulsion study group. Anesthesiology 1999;91:1195–208.
34. Spahn DR, Waschke KF, Standl T, Motsch J, Van Huynegem L, Welte M, Gombotz H, Coriat P, Verkh L, Faithfull S, Keipert P. Use of perflubron emulsion to decrease allogeneic blood transfusion in high-blood-loss non-cardiac surgery: results of a European phase 3 study. Anesthesiology 2002;97:1338–1349.
35. Tremper KK. Perfluorochemical "red blood cell substitutes": the continued search for an indication. Anesthesiology 2002;97:1333–1334.
36. Wahr JA, Levy JH, Kindscher J. Hemodynamic effects of a bovine based oxygen carrying solution in surgical patients. Anesthesiology 1996;85:A347.
37. Creteur J, Vincent JL. Hemoglobin solutions: an "all-in-one" therapeutic strategy in sepsis? Crit Care Med 2000;28: 894–896.
38. Reah G, Bodenham AR, Mallick A, Daily EK, Przybelski RJ. Initial evaluation of diaspirin cross-linked hemoglobin (DCLHB) as a vasopressor in critically ill patients. Crit Care Med 1997;25:1480–1488.
39. Robalino BD, Marwick T, Lafont A, Vaska K, Whitlow PL. Protection against ischemia during prolonged balloon inflation by distal coronary perfusion with use of an autoperfusion catheter or Fluosol. J Am Coll Cardiol 1992;20:1378–1384.
40. Kent KM, Cleman MW, Cowley MJ, Forman MB, Jaffe CC, Kaplan M, King SB 3rd, Krucoff MW, Lassar T, McAuley B. et al. Reduction of myocardial ischemia during percutaneous transluminal coronary angioplasty with oxygenated Fluosol. Am J Cardiol 1990;66:279–284.
41. Creteur J, Sibbald W, Vincent JL. Hemoglobin solutions–not just red blood cell substitutes. Crit Care Med 2000;28: 3025–3034.
42. Teicher BA, Schwartz GN, Dupuis NP, Kusomoto T, Liu M, Liu F, Northey D. Oxygenation of human tumor xenografts in nude mice by a perfluorochemical emulsion and carbogen breathing. Artif Cells, Blood Sub Immobil Biotechnol 1994; 22:1369–1375.
43. Teicher BA. An overview on oxygen carriers in cancer therapy. Artif Cells, Blood Sub Immobil Biotechnol 1995; 23:395–405.
44. Martin SM, Laks H, Drinkwater DC, Stein DG, Capouya ER, Pearl JM, Barthel SW, Chang P, Kaczer E, Bhuta S. Perfluorochemical reperfusion yields improved myocardial recovery after global ischemia. Ann Thorac Surg 1993;55: 954–960.
45. Kloner RA, Hale S. Cardiovascular applications of fluorocarbons in regional ischemia/reperfusion. Artif Cells Blood Substit Immobil Biotechnol 1992;22:1069–1081.
46. Segel LD, Follette DM, Iguidbashian JP, Contino JP, Castellanos LM, Berkoff HA, Kaufman RJ, Schweighardt FK. Posttransplantation function of hearts preserved with fluorochemical emulsion. J Heart Lung Trans 1994;13: 669–680.
47. Grunert A, Qiu H, Muller I, Schuh S, Steinbach G, Wennauer R, Wolf C, Von Schenck H. A new extracorporeal perfusion system: prolongation of liver organ vitality beyond 24 hours. Ann N Y Acad Sci 1994;723:488–490.
48. Premaratne S, Harada RN, Chun P, Suehiro A, McNamara JJ. Effects of perfluorocarbon exchange transfusion on reducing myocardial infarct size in a primate model of ischemia-reperfusion injury: a prospective, randomized study. Surgery 1995;117:670–676.
49. Cole DJ, Schell RM, Drummond JC, Pryzbelski RJ, Marcantonio S. Focal cerebral ischemia in rats: effect of hemodilution with alpha-alpha cross-linked hemoglobin on brain injury and edema. Can J Neurol Sci 1993;20:30–36.
50. Gauger PG, Overbeck MC, Chambers SD, Cailipan CI, Hirschl RB. Partial liquid ventilation improves gas exchange and increases EELV in acute lung injury. J Appl Physiol 1998;84:1566–1572.
51. Hirschl RB, Pranikoff T, Wise C, Overbeck MC, Gauger P, Schreiner RJ, Dechert R, Bartlett RH. Initial experience with partial liquid ventilation in adult patients with the acute respiratory distress syndrome. JAMA 1996;275:383–389.
52. Wong DH. Liquid ventilation: more than "PEEP in a bottle"? Crit Care Med 1999;27:1052–1053.
53. Colton DM, Till GO, Johnson KJ, Dean SB, Bartlett RH, Hirschl RB. Neutrophil accumulation is reduced during partial liquid ventilation. Crit Care Med 1998;26:1716–1724.
54. Mrozek JD, Smith KM, Bing DR, Meyers PA, Simonton SC, Connett JE, Mammel MC. Exogenous surfactant and partial liquid ventilation: physiologic and pathologic effects. Am J Resp Crit Care Med 1997;156:1058–1065.
55. Hirschl RB, Conrad S, Kaiser R, Zwischenberger JB, Bartlett RH, Booth F, Cardenas V. Partial liquid ventilation in adult patients with ARDS: a multicenter phase I-II trial. Adult PLV Study Group. Ann Surg 1998;228:692–700.
56. Rabiner SF, Friedman LH. The role of intravascular haemolysis and the reticulo-endothelial system in the production of a hypercoagulable state. Br J Haematol 1968;14: 105–118.
57. Kasper SM, Grune F, Walter M, Amr N, Erasmi H, Buzello W. The effects of increased doses of bovine hemoglobin on hemodynamics and oxygen transport in patients undergoing preoperative hemodilution for elective abdominal aortic surgery. Anesth Analg 1998;87:284–291.
58. Standl T, Burmeister MA, Horn EP, Wilhelm S, Knoefel WT, Schulte am Esch J. Bovine haemoglobin-based oxygen

carrier for patients undergoing haemodilution before liver resection. Br J Anaesth 1998;80:189–194.
59. Johnson JL, Moore EE, Offner PJ, Haenel JB, Hides GA, Tamura DY. Resuscitation of the injured patient with polymerized stroma-free hemoglobin does not produce systemic or pulmonary hypertension. Am J Surg 1998;176:612–617.
60. Gould SA, Moore EE, Hoyt DB, Burch JM, Haenel JB, Garcia J, DeWoskin R, Moss GS. The first randomized trial of human polymerized hemoglobin as a blood substitute in acute trauma and emergent surgery. J Am Coll Surg 187:113–20; discussion 1998; 120–122.
61. Sehgal LR, Rosen AL, Gould SA, Sehgal HL, Moss GS. Preparation and in vitro characteristics of polymerized pyridoxylated hemoglobin. Transfusion (Paris) 1983;23: 158–162.
62. Privalle C, Talarico T, Keng T, DeAngelo J. Pyridoxalated hemoglobin polyoxyethylene: a nitric oxide scavenger with antioxidant activity for the treatment of nitric oxide-induced shock. Free Rad Biol Med 2000;28:1507–1517.
63. DeAngeles DA, Scott AM, McGrath AM, Korent VA, Rodenkirch LA, Conhaim RL, Harms BA. Resuscitation from hemorrhagic shock with diaspirin cross-linked hemoglobin, blood, or hetastarch. J Trauma-Injury Inf Crit Care 42:406–412; discussion 1997; 412–414.
64. Saxena R, Wijnhoud AD, Carton H, Hacke W, Kaste M, Przybelski RJ, Stern KN, Koudstaal PJ. Controlled safety study of a hemoglobin-based oxygen carrier, DCLHB, in acute ischemic stroke. Stroke 1999;30:993–996.
65. Rudolph AS. Encapsulated hemoglobin: current issues and future goals. Artif Cells, Blood Sub Immobiliz Biotechnol 1994;22:347–360.
66. Hughes GS Jr., Antal EJ, Locker PK, Francom SF, Adams WJ, Jacobs EE Jr Physiology and pharmacokinetics of a novel hemoglobin-based oxygen carrier in humans. Crit Care Med 1996;24:756–764.
67. Winslow RM. Blood substitutes. Adv Drug Del Rev 2000;40: 131–142.
68. USDA Interim Rule on Import Restrictions of Ruminant Material from Europe. Fed Proc 1998;63:406–408.
69. Ulatowski JA, Nishikawa T, Matheson-Urbaitis B, Bucci E, Traystman RJ, Koehler RC. Regional blood flow alterations after bovine fumaryl beta beta-crosslinked hemoglobin transfusion and nitric oxide synthase inhibition. Crit Care Med 1996;24:558–565.
70. Ning J, Wong LT, Christoff B, Carmichael FJ, Biro GP. Haemodynamic response following a 10% topload infusion of Hemolink™ in conscious, anaesthetized and treated spontaneously hypertensive rats. Transfus Med 2000;10:13–22.
71. Krieter H, Hagen G, Waschke KF, Kohler A, Wenneis B, Bruckner UB, van Ackern K. Isovolemic hemodilution with a bovine hemoglobin-based oxygen carrier: effects on hemodynamics and oxygen transport in comparison with a non-oxygen-carrying volume substitute.[See comment]. J Cardiothor Vas Anesthes 1997;11:3–9.
72. Maxwell RA, Gibson JB, Fabian TC, Proctor KG. Resuscitation of severe chest trauma with four different hemoglobin-based oxygen-carrying solutions. J Trauma-Injury Inf Crit Care 49:200–209; discussion 2000; 209–211.
73. Standl T, Wilhelm S, Horn EP, Burmeister M, Gundlach M, Schulte am Esch J. Preoperative hemodilution with bovine hemoglobin. Acute hemodynamic effects in liver surgery patients. Anaesthesist 1997;46:763–770.
74. Cheng DC, Ralph-Edwards A, Mazer CD, Carmichael FJL, Biro GP. The hemodynamic effects of the red cell substitute Hemolink™ (o-rafinose cross-linked human hemoglobin) on vital signs in patients undergoing CABG surgery. Anesthesiology 2000;93:A-180.
75. LaMuraglia GM, O'Hara PJ, Baker WH, Naslund TC, Norris EJ, Li J, Vandermeersch E. The reduction of the allogenic transfusion requirement in aortic surgery with a hemoglobin-based solution. J Vasc Surg 2000;31:299–308.
76. Schubert A, Mascha E, O'Hara JF. Synthetic hemoglobin reduces perioperative blood transfusions in vascular, orthopedic and abdominal surgery. Anesthesiology 2000;93:180.
77. Lamy ML, Daily EK, Brichant JF, Larbuisson RP, Demeyere RH, Vandermeersch EA, Lehot JJ, Parsloe MR, Berridge JC, Sinclair CJ, Baron JF, Przybelski RJ. Randomized trial of diaspirin cross-linked hemoglobin solution as an alternative to blood transfusion after cardiac surgery. The DCLHB cardiac surgery trial collaborative group. Anesthesiology 2000;92:646–656.
78. Carmichael FJ, Ali AC, Campbell JA, Langlois SF, Biro GP, Willan AR, Pierce CH, Greenburg AG. A phase I study of oxidized raffinose cross-linked human hemoglobin. Crit Care Med 2000;28:2283–2292.
79. Patel RP. Biochemical aspects of the reaction of hemoglobin and no: implications for hb-based blood substitutes. Free Rad Biol Med 2000;28:1518–1525.
80. Kim HW, Greenberg AG. Ferrous sulphate scavenging of endothelium derived nitric oxide is a principal mechanism for hemoglobin mediated vasoactivities in isolated rat thoracic aorta. Artf Cells, Blood Sub Immobil Biotechnol 1997;25:121–133.
81. Su D, Roth RI, Levin J. Hemoglobin infusion augments the tumor necrosis factor response to bacterial endotoxin (lipopolysaccharide) in mice. Crit Care Med 1999;27:771–778.
82. Freilich E, Freilich D, Hacker M, Leach L, Patel S, Hebert J. The hemodynamic effects of diaspirin cross-linked hemoglobin in dopamine-resistant endotoxic shock in swine. Art Cells, Blood Sub Immobiliz Biotechnol 2002;30:83–98.
83. Zimmerman JJ. Deciphering the dark side of free hemoglobin in sepsis. Crit Care Med 1999;27:685–686.
84. Winslow RM. Current status of blood substitute research: towards a new paradigm. J Intern Med 2003;253:508–517.
85. Kramer GC. Counterintuitive red blood cell substitute–polyethylene glycol-modified human hemoglobin. Crit Care Med 2003;31:1882–1884.
86. Wettstein R, Tsai AG, Erni D, Winslow RM, Intaglietta M. Resuscitation with polyethylene glycol-modified human hemoglobin improves microcirculatory blood flow and tissue oxygenation after hemorrhagic shock in awake hamsters. Crit Care Med 2003;31:1824–1830.
87. Slanetz PJ, Lee R, Page R, Jacobs EE Jr, LaRaia PJ, Vlahakes GJ. Hemoglobin blood substitutes in extended preoperative autologous blood donation: an experimental study. Surgery 1994;115:246–254.
88. Lee R, Neya K, Svizzero TA, Vlahakes GJ. Limitations of the efficacy of hemoglobin-based oxygen-carrying solutions. J Appl Physiol 1995;79:236–242.
89. Vlahakes GJ. Hemoglobin solutions come of age. Anesthesiology 2000;92:637–638.
90. Levy JH. Hemoglobin-based oxygen-carrying solutions: close but still so far. Anesthesiology 2000;92:639–641.
91. Hughes GS Jr, Francome SF, Antal EJ, Adams WJ, Locker PK, Yancey EP, Jacobs EE Jr. Hematologic effects of a novel hemoglobin-based oxygen carrier in normal male and female subjects. J Lab Clin Med 1995;126:444–451.
92. Schumacher YO, Ashenden M. Doping with artificial oxygen carriers: an update. Sports Med 2004;34:141–150.
93. Levy JH, Goodnough LT, Greilich PE, Parr GV, Stewart RW, Gratz I, Wahr J, Williams J, Comunale ME, Doblar D, Silvay G, Cohen M, Jahr JS, Vlahakes GJ. Polymerized bovine hemoglobin solution as a replacement for allogeneic red blood cell transfusion after cardiac surgery:

results of a randomized, double-blind trial. J Thor Cardiovas Sur 2002;124:35–42.
94. Gonzalez P, Hackney AC, Jones S, Strayhorn D, Hoffman EB, Hughes G, Jacobs EE, Orringer EP. A phase I/II study of polymerized bovine hemoglobin in adult patients with sickle cell disease not in crisis at the time of study. J Investig Med 1997;45:258–264.
95. Feola M, Simoni J, Angelillo R, Luhruma Z, Kabakele M, Manzombi M, Kaluila M. Clinical trial of a hemoglobin based blood substitute in patients with sickle cell anemia. Surg Gynecol Obs 1992;174:379–386.
96. Mullon J, Giacoppe G, Clagett C, McCune D, Dillard T. Transfusions of polymerized bovine hemoglobin in a patient with severe autoimmune hemolytic anemia. N Engl J Med 2000;342:1638–1643.
97. Sloan EP. The clinical trials of diaspirin cross-linked hemoglobin (DCLHB) in severe traumatic hemorrhagic shock: the tale of two continents. Int Care Med 2003;29: 347–349.
98. Gould SA, Moore EE, Hoyt DB, Ness PM, Norris EJ, Carson JL, Hides GA, Freeman IH, DeWoskin R, Moss GS. The life-sustaining capacity of human polymerized hemoglobin when red cells might be unavailable. J Am Coll Surg 195: 445–452; discussion 2002; 452–455.
99. Lieberthal W, Fuhro R, Freedman JE, Toolan G, Loscalzo J, Valeri CR. O-raffinose cross-linking markedly reduces systemic and renal vasoconstrictor effects of unmodified human hemoglobin. J Pharmacol Exper Therap 1999;288:1278–1287.
100. Cheng DC, Mazer CD, Martineau R, Ralph-Edwards A, Karski J, Robblee J, Finegan B, Hall RI, Latimer R, Vuylsteke A. A phase ii dose-response study of hemoglobin raffimer (Hemolink) in elective coronary artery bypass surgery. J Thor Cardiov Sur 2004;127:79–86.
101. Greenburg AG, Kim HW, Hemolink Study Group. Use of an oxygen therapeutic as an adjunct to intraoperative autologous donation to reduce transfusion requirements in patients undergoing coronary artery bypass graft surgery. J Am Coll Surg 198:373–383; discussion 2004; 384–385.
102. Conover CD, Lejeune L, Shum K, Gilbert C, Shorr RG. Physiological effect of polyethylene glycol conjugation on stroma-free bovine hemoglobin in the conscious dog after partial exchange transfusion. Artif Organs 1997;21:369–378.
103. Teicher BA, Ara G, Herbst R, Takeuchi H, Keyes S, Northey D. Peg-hemoglobin: effects on tumor oxygenation and response to chemotherapy. In Vivo 1997;11:301–311.
104. Linberg R, Conover CD, Shum KL, Shorr RG. Increased tissue oxygenation and enhanced radiation sensitivity of solid tumors in rodents following polyethylene glycol conjugated bovine hemoglobin administration. In Vivo 1998;12: 167–173.
105. Serna DL, Powell LL, Kahwaji C, Wallace WC, West J, Cogert G, Smulowitz P, Steward E, Purdy RE, Milliken JC. Cardiac function after eight hour storage by using polyethylene glycol hemoglobin versus crystalloid perfusion. ASAIO J 2000;46:547–552.
106. Meng FT, Zhang WZ, Ma GH, Su ZG. The preparation and characterization of monomethoxypoly(ethylene glycol)-*b*-poly-*dl*-lactide microcapsules containing bovine hemoglobin. Artf Cells, Blood Sub Immobil Biotechnol 2003;31:279–292.
107. Phillips WT, Klipper RW, Awasthi VD, Rudolph AS, Cliff R, Kwasiborski V, Goins BA. Polyethylene glycol-modified liposome-encapsulated hemoglobin: a long circulating red cell substitute. J Pharmacol Exp Ther 1999;288:665–670.

See also BIOCOMPATIBILITY OF MATERIALS; BLOOD COLLECTION AND PROCESSING; BLOOD GAS MEASUREMENTS.

BONDING, ENAMEL.

See RESIN-BASED COMPOSITES.

BONE AND TEETH, PROPERTIES OF

RODERIC LAKES
University of Wisconsin
Madison, Wisconsin

J. LAWRENCE KATZ
University of Missouri
Kansas City, Missouri

INTRODUCTION

Bone has a variety of functions in the body of which some of the most important are structural in nature: protection of vulnerable body parts, support of the body, and to provide muscle attachments. A knowledge of the mechanical and adaptive properties of bone is useful in the design and use of prostheses that replace a bone or a portion of a bone. Mechanical properties of bone are also of interest in trauma biomechanics and in efforts to prevent injury to the body. As for teeth, they also are replaced by artificial materials which are called upon to perform the mechanical functions of the original tooth. This article contains a survey of known properties of bone and teeth and their components collagen and apatite, with an emphasis on bone and its mechanical properties.

A voluminous literature is available dealing with the properties of bone and to a lesser extent of teeth's major constituents: collagen and apatite. Reported properties are often found to differ. Some of the differences arise from the fact that bone and tooth structures are of biological origin and consequently vary depending on the individual and on the part of the body from which the specimen is taken. Other differences are due to experimental technique and variations in environmental conditions during experiments. Of necessity, results presented in this article are selected from a large mass of published reports. The authors have endeavored to select results obtained by good techniques and representative of accepted values. Nevertheless, other results obtained by equally good techniques may be expected to differ somewhat as a result of biological variability. Therefore, it is suggested that additional measurements of the properties of bone and teeth can be found in several of the source books listed in the Bibliography [see (1–10)].

MECHANICAL PROPERTIES OF COMPACT BONE

Compact Bone Structure

The mechanical properties of bone are inseparably related to its structure. Bone tissue is a complex composite material that at different levels of scale exhibits fibrous, porous, and particulate microstructural features (1–5). The following constituents are present in bone: mineral, protein, other organic materials, and fluids such as water. The mineral, principally a carbonated apatite, where the

carbonate group substitutes in part for the phosphate group [$Ca_{10}(PO_4,CO_3)_6(OH)_2$] (5,6). In mature bovine cortical bone (similar to human cortical bone) it occurs as microcrystalline inclusions of plate-like shape (mineralites) of dimensions $\sim 0.7 \times 11 \times 17$ nm (11). However, in young postnatal bovine bone, the mineralites are thicker, shorter, and narrower, [i.e. $2 \times 6 \times 9$ nm (12)]. These mineralite sizes measured by AFM are closest to the actual values as, "AFM yields the full three-dimensional structure of mineralites rather than a projection $\cdots$" thus providing the full shape of each mineralite measured. The other major techniques for measuring mineralite sizes, transmission electron microscopy (TEM) and X-ray diffraction line broadening, each suffer from artifacts that result in increased sizes of some of the dimensions (3).

On the ultrastructural level (nanoscale), the mineral crystallites are in intimate apposition with fibrils of the protein collagen. A, "(d)iagrammatic depiction of the supramolecular packing of collagen molecules in a fibril $\cdots$" plus the possible arrangement of the mineralites within a fibril is given as Fig. 7 in Ref. 12. These fibrils are from 20 to 200 nm in diameter and are organized into fibers that are in turn arranged in bone into lamellae or layers. The fibers in each lamella run longitudinally, spirally, or nearly circumferentially. Moreover, the orientation of these layers are different in alternate lamellae. A micromechanical model for the Young's modulus of bone has been proposed, based on these histological features, however, it has not yet been tested experimentally. The organization of collagen fibrils differs in woven bone and lamellar bone. Mineralized collagen fibrils and isolated crystals from the mid-diaphyses of human fetal femurs were observed with scanning, TEM, and high resolution electron microscopy. The apatite crystals in woven bone are also platelet shaped, similar to mature crystals from lamellar bone. Average crystal dimensions are considerably smaller in woven bone than those of mature crystals in lamellar bone. In diseased bone such as that affected by osteogenesis imperfecta, the apatitic crystals occur in various sizes and shapes; they are oriented and aligned with respect to collagen in a manner that differs from that found in normal calcified tissues.

In compact cortical bone, the lamellae are layers arranged circumferentially around a central canal and form the Haversian system (secondary osteon). Haversian canals typically contain small blood vessels. Haversian bone occurs in the cortices of bones in adult humans (1–3,5) and in the bones of various large animals (1,2). Osteons are roughly cylindrical structures up to $\sim$ 200 μm in diameter. In a long bone, they tend to run approximately parallel to each other and to the bone axis. Figure 1 displays the typical structure of human cortical bone, whereas Fig. 2 displays the typical structure of human cancellous (also known as trabecular or spongy) bone. The large circular or elliptical features in the former figure are the cross-sections of osteons, the concentric layers are lamellae, and the central dark circles are the cross-sections of Haversian canals. The numerous spots among the lamellae are the lacunae, in which the osteocytes, or bone cells live. The lacunae are roughly ellipsoidal and have dimensions of $\sim 10 \times 15 \times 25$ μm. The osteocytes have many thin processes that occupy channels in the bone matrix known as canaliculi; they are too small to be visible in Fig. 1. The mechanical properties of bone depend on its degree of bulk and surface hydration and the details of its structure that depend on variables such as the age, state of health, and level of physical activity of the individual, the location of the bone in the body, as well as the rate and direction at which load is applied to the bone.

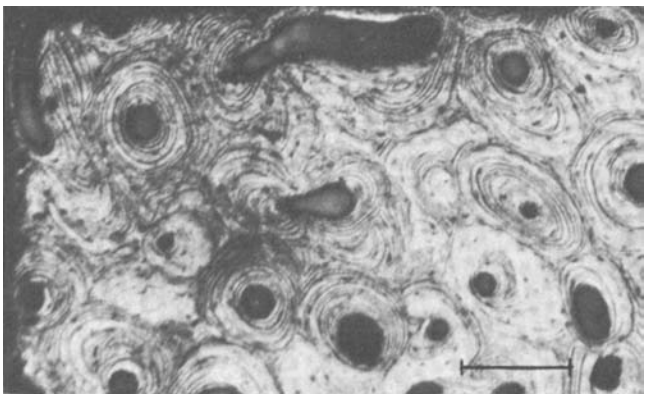


Figure 1. Human Haversian bone. Reflected light micrograph of a specimen cut perpendicular to the bone axis. Scale mark: 200 μm.

The emphasis in this article is on wet skeletal tissues, bone and teeth, from healthy adults, with occasional references to dry tissues for comparison. However, there are many studies of mammalian skeletal tissues as they provide a useful counterpart to the human studies; studies of the properties of bovine bone and teeth are the most numerous in this respect. The structures of both mature bovine cortical and cancellous bone are very similar to those corresponding human tissues Figs. 1 and 2, respectively. Young bovine cortical bone structure is quite different as seen in Fig. 3. This plexiform (or lamellar) bone resorbs and remodels to Haversian bone as the animal matures. Comparison of the young and mature bovine bone properties provides added insight into the importance of structure in determining the mechanical properties of bone. Thus, where appropriate, properties of bovine bone and teeth are included in Tables 1–7.

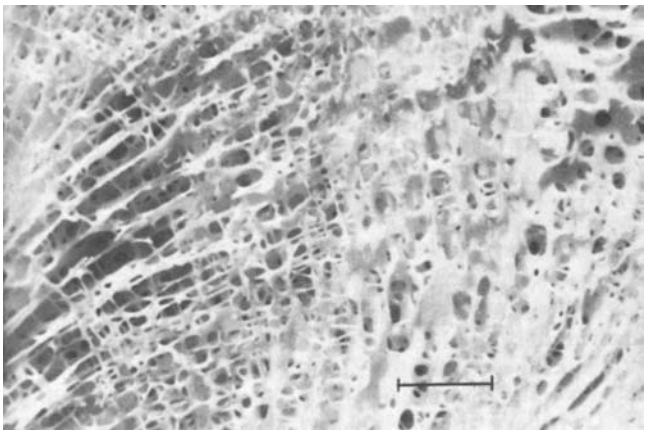


Figure 2. Human cancellous bone from the proximal femur. The marrow has been removed. Scale mark: 5 mm.

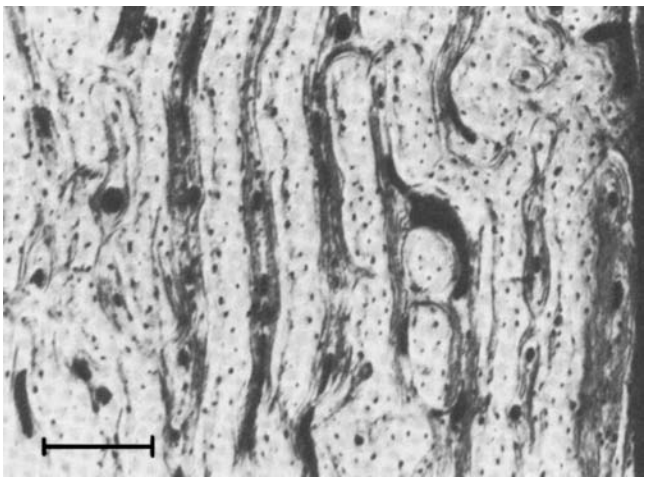


Figure 3. Bovine plexiform bone. Reflected light micrograph of a specimen cut perpendicular to the bone axis. Scale mark: 200 μm.

Elasticity of Compact Bone: Comparison with Teeth and Other Materials

Elastic materials deform under load; elasticity entails reversible behavior in which the material returns to its original configuration after the load is removed. If the load is sufficiently small, there is a linear relationship between stress and strain. For simple tension or compression, the constant of proportionality is referred to as Young's modulus E and for shear or torsion it is the shear modulus G(4,29). The stiffness of human compact bone as quantified by Young's modulus typically lies between 12 and 25 GPa for tension or compression depending on sample orientation and experimental technique (17), (1 GPa = 145,000 lb·in.$^{-2}$). It has been suggested that the tensile and compressive elastic moduli are not equal, but at small strain, the best evidence indicates that tensile and compressive properties are identical (13).

In Table 1, some examples of the elastic moduli of wet human bone and teeth are compared with those of various other materials. The mechanical testing strain rates for bone are faster than those encountered by bones in walking, but are perhaps comparable to those in vigorous activities. They are slower than those that occur during fracture. Compact bone is ∼ 10 times stiffer than most "rigid" polymers, but is about one-tenth as stiff as common metals such as steel. Table 1 also shows the relationship between stiffness and density. This relationship governs the structural efficiency of whole bones. In view of the fact that the skeleton represents ∼ 17% of the weight of the human body, structural efficiency of bones is relevant to their performance in the body. For example, the overall rigidity per unit weight of a bone acting as a short column or as a tensile member is proportional to the modulus to density ratio, E/ρ, of the bone *tissue* of which it is made. By contrast, for a given weight of material, the Euler buckling load of a bone acting as a slender column or the bending stiffness of a bone acting as a beam of constant shape is proportional to E/ρ^2(2,6). Density enters these relations in a different way since the rigidity of a short column depends on its cross-sectional area while the bending rigidity of a beam is governed by its moment of inertia.

A more complete listing of the properties of both human and bovine bone and teeth is given in Table 2, the latter data are included because of the similarity in hierarchical structure and organization of mature bovine compact bone with that of human compact bone. These data were obtained using various techniques, mechanical testing, ultrasonic wave propagation (UWP). In the low megahertz (MHz) region (2–5 MHz), scanning acoustic microscopy (SAM) in the high megahertz region (400–600 MHz), and nanoindentation. The first three techniques also were used to obtain the teeth data. Here too, the properties are strongly dependent on the sample location and orientation, for example, the range in dentin properties over the sample surface obtained by SAM (21), Table 2. Corresponding properties of human trabecular bone volumes and individual trabeculae are given in Table 7.

Because SAM is not as well documented compared to either UWP or mechanical testing (MT) in obtaining elastic properties, a description of the technique follows below.

Scanning Acoustic Microscopy

The development of SAM (30,31); see also (9) enabled the analysis of the biomechanical properties of materials at much higher resolution than was previously achieved using traditional ultrasonic wave propagation techniques.

Table 1. Bone, Teeth, and Other Materials: Stiffness

Material	Young's Modulus E, GPa	Density (ρ), g·cm^{-3}	E/ρ	E/ρ^2
Human compact bone				
longitudinal direction (13)	17	1.8	9.4	5.2
transverse direction	12.5			
Tooth dentin (7,14)	18	2.1	8.6	4.1
Tooth enamel (7,15)	50	2.9	17	6.0
Polyethylene (high density) (4)	0.5	0.95	0.53	0.55
Polymethyl methacrylate (4)	3.0	1.2	2.5	2.2
Steel(structural)	200	7.9	25	3.2
Aluminum	70	2.7	26	9.5
Granite	70	2.8	25	9.1
Concrete	25	2.3	11	4.6
Wood(pine)	11	0.6	18	30

Table 2. Elastic Properties of Wet Human and Bovine Bone and Teeth

Material	Young's Modulus, GPa	References
Human compact bone		
axial direction (femur)	27.7 (U)[a]	16
	17.6 (M)[a]	17
	23.4 (S)[a]	18
(osteons)	22.4 (N)[a]	19
(interstitial lamellae)	25.7 (N)	19
radial direction (femur)	18.9 (U)	16
	12.5 (M)	17
	13.0 (N)	19
Human teeth		
dentin	13.0 (S)	20
	16.4–38.6 (S)	21
enamel	62.7 (S)	20
Bovine compact bone		
axial direction (tibia)	36.0 (M)	2
axial direction (femur)	22.7 (M)	17
axial direction (femur)	21.9 (U)	22
radial direction (tibia)	22.8 (M)	2
radial direction (femur)	10.3 (M)	17
radial direction (femur, average)	13.1 (U)	22
Bovine teeth		
dentin	26.3 (U)	23,24
dentin	25.2 (U)[b]	25
enamel	105.1 (U)	23,24
enamel	97.8 (U)[b]	25

[a] U = Ultrasonics; M = Mechanical Testing; S = Scanning Acoustic Microscopy; N = Nanoindentation.

[b] Average of measurements of several samples.

Table 5. Comparison of Materials: Tensile Strength

Material	Strength σ_{ult}, MPa	Density ρ, g · cm^{-3}	σ_{ult}/ρ
Human femoral compact bone (17),			
longitudinal direction	148	2.0	74
transverse direction	49	2.0	25
Bovine femoral plexiform bone, (13)			
longitudinal direction	167	2.0	83
Tooth dentin (8), average	271	2.1	130
Tooth enamel (8), average	275	2.9	95
Polyethylene (high density)	20–40	0.95	21–42
Poly(methyl methacrylate) (PMMA)	70	1.2	59
Steel(structural)	400	7.8	51
Aluminum(1100-H14)	110	2.7	41
Granite	20	2.8	7.2
Concrete(compression)	28	2.3	12

A significant advantage of SAM is the ability to investigate the properties of internal and subsurface structures in addition to the surface properties of most materials, including those that are optically opaque. Another advantage of special interest for studying biological materials is that a liquid couplant must be used to transmit the acoustic waves from the acoustic lens to the specimen being studied, thus the specimen is kept wet during all measurements. Therefore, fresh tissue specimens can be used as well as embedded specimens. In addition, the use of high quality,

Table 3. Elastic Anisotropy of Bovine and Human Bone

	Tensorial Elastic Moduli (GPa), Determined Ultrasonically Wet Bovine Femur (26)		Dry Human Femur (27)
Elastic constant	Haversian (transverse isotropic)	Plexiform (orthotropic)	Haversian (transverse isotropic)
C_{11}	21.2	22.4	23.4
C_{22}	21.0	25.0	
C_{33}	29.0	35.0	32.5
C_{44}	6.30	8.20	8.71
C_{55}	6.30	7.10	
C_{66}	5.40	6.10	
C_{12}	11.7	14.0	9.06
C_{13}	12.7	15.8	
C_{23}	11.1	13.6	9.11

Table 4. Elastic Anisotropy of Bone[a]

Young's Moduli, GPa		Shear Moduli, GPa		Poisson's Ratios, Dimensionless	
Human	Bovine	Human	Bovine	Human	Bovine
$E = 17.0$[b]	22	$G = 3.6$	5.3	$v = 0.58$	0.30
$E = 11.5$[c]	15	$G = 3.3$	6.3	$v = 0.31$	0.11
$E = 11.5$[d]	12	$G = 3.3$	7.0	$v = 0.31$	0.21

[a] Technical elastic moduli. Wet human femoral bone by mechanical testing (13) and bovine femoral bone by ultrasound (23).

[b] Radial direction.

[c] Circumferential direction.

[d] Longitudinal direction.

Table 6. Elastic Properties of Apatites

Material	Young's modulus, GPa	Bulk modulus, GPa	Shear modulus, GPa	Reference
Hydroxyapatite				
(polycrystalline)	117	88.0	45.5	27
(single crystal, modeling[a])	120	111	45.3	8
Fluoroapatite				
(polycrystalline)	120	94.0	46.4	27
(single crystal, ultrasonics[a])	130	117	49.4	28
Chloroapatite				
(polycrystalline)	94.3	68.5	37.1	27

[a]Calculated from single-crystal elastic constants.

high frequency focusing acoustic lenses permits examination of the elastic properties of biological materials on a microscope scale comparable to the optical histology studies. The heart of the SAM is a spherical lens formed at the interface between a high acoustic velocity solid (e.g., sapphire) and the low acoustic velocity couplant liquid. Due to the high acoustic "refractive index", spherical aberration is negligible and an acoustic beam displaying significant convergence can be obtained. A radio frequency (RF) signal is generated from the transmitter. The acoustic lens is equipped with a piezoelectric transducer that converts the RF signal into an acoustic wave. This signal is then made to converge by the lens and propagates to the specimen through the coupling liquid. When reaching the surface of the specimen, a part of the acoustic wave is reflected back through the lens to the transducer, that, acting now as a receiver, transforms the acoustic signal into an RF signal. The amplitude of the echo reflected back to the lens is a measure of the acoustic reflectivity of the surface of the investigated material at the point in focus. It is proportional to the reflection coefficient, r, which is related to the acoustic impedances of the liquid couplant, Z_1, and the investigated material, Z_2, by the equation r given in Fig. 4. Acoustic impedance, Z, is measured in Rayls and is defined as $Z = \rho v$, where ρ is the material density and v is the velocity of the longitudinal (dilatational) acoustic wave propagating in the direction perpendicular to the surface. The images present the variations in acoustic signals that originate either from the intrinsic acoustic reflectivity of the material surface or through interference occurring between different surface and subsurface reflected signals. In the former case, surface imaging, the acoustic reflectivity variations are a result of the local changes in acoustic impedance (32).

Table 7. Elastic Modulus of Trabecular Bone Volumes and Trabeculae

Trabeculae	Elastic Modulus, GPa	Reference
Human trabecular bone (Trabeculae)	17.4 (S)	18
(longitudinal)	19.4 (N)	19
(transverse)	15.0 (N)	19
Human trabecular bone volumes		
(proximal tibia)	0.445 (M)	3

Figure 5 is a plot of elastic stiffness (GPa) versus reflection coefficient, r, for a wide range of materials. Bone has an acoustic impedance in the neighborhood of $Z = 7.5$ Mrayls, yielding a reflection coefficient in the neighborhood of $r = 0.67$. Of course, Z for bone can vary over a considerable range depending on a number of factors that would affect both the density and structural cohesivity of the specific specimen being measured. Likewise, Z for water will vary depending on the temperature at which the couplant fluid is maintained during the experiment, for example, at 0 °C, $Z(H_2O) = 1.40$ Mrayl; at body temperature, 37 °C, $Z(H_2O) = 1.51$ Mrayl; and at 60 °C, $Z(H_2O) = 1.53$, due mainly to the variations in water's acoustic properties with temperature.

As described above, the high frequency mode at 400 and 600 MHz also has been used to study the *in vitro* micromechanical elastic properties of human trabecular and

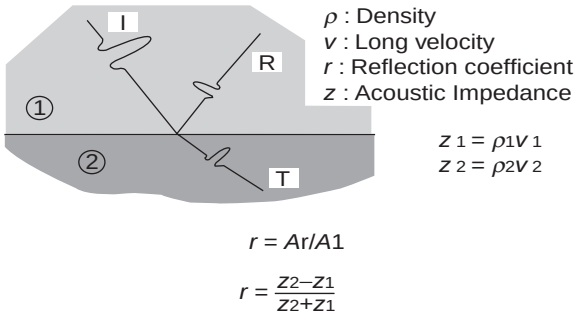


Figure 4. Schematic diagram of the incident, reflected, and transmitted signals at the interface between two materials.

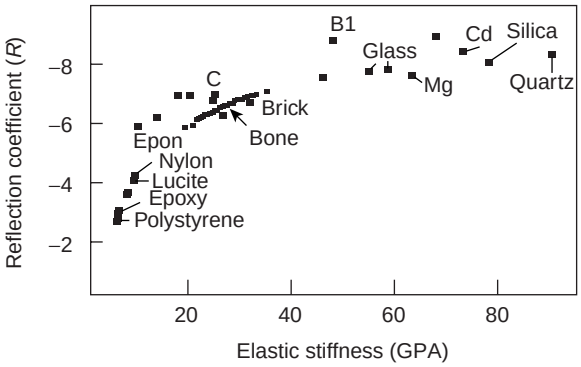


Figure 5. Reflection coefficient, r versus Modulus, E.

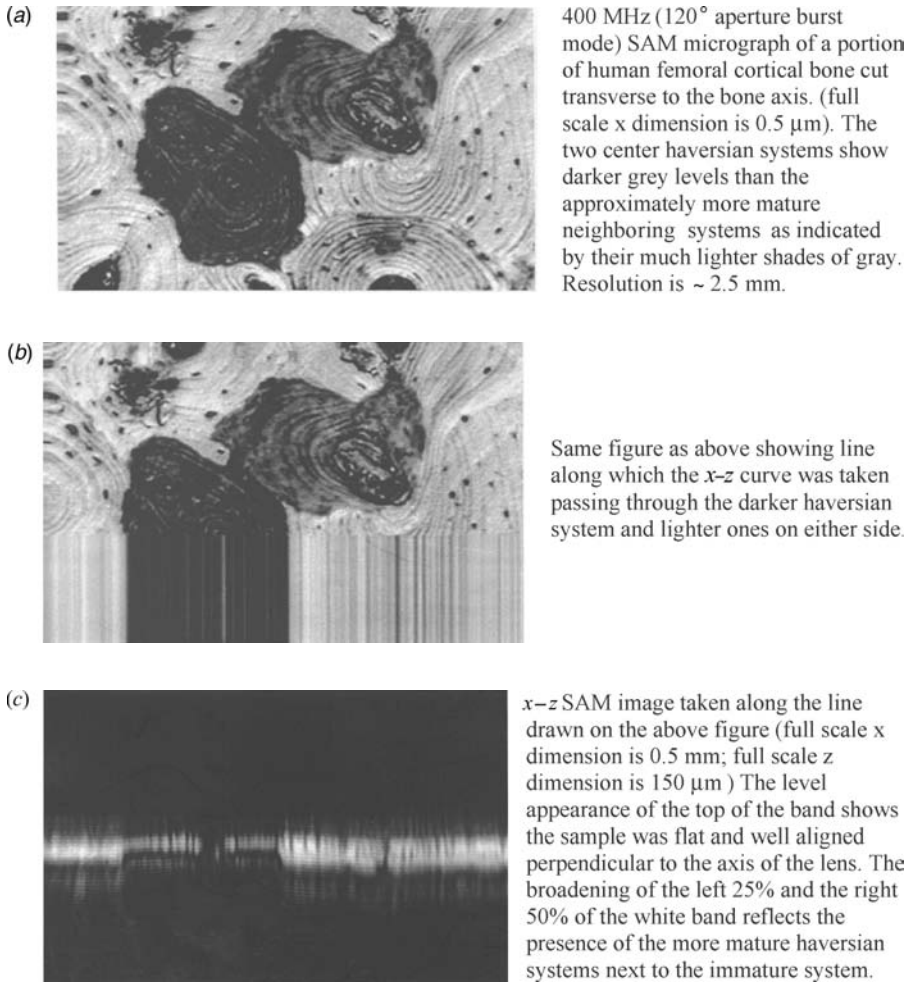


Figure 6. (a) The 400 MHz SAM image of two hypomineralized osteons from human femoral cortical bone. (b) Same area as in part a showing the line along which an x–z interference image was taken. (c) The x–z Interference image taken along the line shown in part b.

compact cortical femoral bone (15,31). In this study, they were able to image the properties of the individual osteonic lamellae at high enough resolution so that three new micromechanical observations were made: (*1*) the outermost lamellae, always appear to be more compliant; (*2*) the outermost lamellae of adjacent abutting osteons appear to have the same acoustic impedance (and thus Young's modulus), even though structurally distinct; and (*3*) adjacent lamellae within an osteon alternate in their acoustic impedance (and thus Young's modulus).

Scanning acoustic microscopy is particularly powerful in providing physical evidence of the possible differences in sound velocity in regions of significant differences in gray level, that is, the darker the gray level, the lower the acoustic impedance, the brighter the gray level, the higher the acoustic impedance, Fig. 6a–c. Figure 6a is a SAM micrograph (400 MHz Burst mode, 120° aperture lens) of human femoral cortical bone (18). The two much darker Haversian systems in the middle of the image are surrounded by the types of Haversian systems (secondary osteons) usually observed in normal bone. The startling difference in gray levels could arise due to out-of-focus artifacts developed in cutting or polishing the specimen. However, performing a x–z curve along the line depicted in Fig. 6b, which goes through the lower, darker Haversian system and the surrounding tissues, yields the image, Fig. 6c; note that the upper level of the broad band is essentially level, indicating that the specimen surface is everywhere level and normal to the acoustic beam. More important, the narrow band and secondary reflection corresponding to the dark Haversian reflects the lower Z and r for the Haversian. A possible explanation for the Haversians is that there was an arrested state of development leading possibly to a hypomineralized area. Unfortunately, we do not have information concerning the possibility of a disease state or a drug modality responsible for these two underdeveloped Haversians that possibly formed just prior to the death of the individual. They are illustrated here to show the ability of the SAM to permit highly sensitive measurements, at high resolution, of variations in Z and r due to remodeling.

In order to convert from reflection coefficient to Young's modulus on the SAM scans of materials of unknown properties, taken with the Olympus UH3 SAM, it is necessary to develop three calibration curves—voltage (V) versus reflection coefficient (r); reflection coefficient (r) versus acoustic impedance (Z); and acoustic impedance (Z) versus Young's modulus (E)—based on using the values of known materials ranging from polymers at the low end of r to metals and ceramics at the high end (18).

The stiffness of compact bone tissue depends on the bone from which it is taken. Fibular bone has a Young's modulus $\sim 18\%$ greater, and tibial bone $\sim 7\%$ greater than that of femoral bone (33). The differences are associated with differences in the histology of the bone tissue. Femoral bone, for example, has a high proportion of osteons that appear light in the polarizing microscope and that have a preponderantly circumferential collagen fiber orientation (34). The histology of a bone is unquestionably related to its function in the body. The relationship has been explored in some detail in the context of bones of very different function in various species of animals (1,2).

Bone is elastically anisotropic, that is, its properties depend on direction. Such behavior is unlike that of steel, aluminum, and most plastics, but is similar to that of wood. Anisotropic properties of bone are shown in Tables 3 and 4. As can be seen from the data in Table 4, human compact femoral bone is ~ 1.5 times as stiff in the longitudinal direction as it is in the transverse directions. The shear modulus G, determined from a torsion test upon a specimen aligned with the bone axis, is ~ 3.3 GPa (13), so that $E/G = 5.15$. Such a small value of G in comparison with Young's modulus E is a further manifestation of the anisotropy of cortical bone. For normal isotropic materials (positive Poisson's ratio), E/G, must lie between 2 and 3 and is typically ~ 2.6. Detailed studies of the anisotropy of bone have been conducted using ultrasonic methods as well as by mechanical testing. The elastic constants given in Table 3 are components of the elastic modulus tensor C_{ijkl} in the following relation between stress σ_{ij} and strain e_{kl} in an anisotropic material. The 3 axis is here assumed to be the bone's longitudinal axis, the 2 direction is circumferential, and the 1 direction is radial.

$$\sigma_{ij} = \sum_{k=1}^{3}\sum_{l=1}^{3} C_{ijkl} e_{kl}$$

In this equation, indexes i and j can have values 1, 2, or 3, so that there are nine components of stress of which six are independent. Since there are six independent components of strain, the number of independent C values is reduced from 81 to 36. Consideration of conservation of mechanical energy further reduces the number of elastic constants to 21, for the least symmetric anisotropic material (7). Materials that have some structural symmetry are described by fewer anisotropic elastic constants. For example, an orthotropic material, (e.g., wood), with three perpendicular planes of symmetry is described by nine constants. A material with transverse isotropic symmetry, (i.e., one that appears the same under an arbirary rotation about an axis), is described by five constants. In crystal physics, the former symmetry is referred to as orthotropic, while the latter is referred to as hexagonal. An isotropic material appears the same under any rotation and has the same properties in any direction. Two independent elastic constants are needed for the description of the elastic behavior of such a material. Equation 1 can be reduced to a more tractable form by a compaction of the counting indices, the so-called Einstein notation, that is, $\sigma_i = C_{ij} e_j$, where $11 \rightarrow 1$, $22 \rightarrow 2$, $33 \rightarrow 3$, $23 \rightarrow 4$, $13 \rightarrow 5$, $12 \rightarrow 6$ (e.g., $C_{1123} \rightarrow C_{14}$). This reduced notation is used in Table 3.

Both structural considerations and experiments indicate that human compact bone has five independent elastic constants and therefore exhibits transverse isotropic symmetry (27,35,36). Results of a different ultrasonic experiment suggest different stiffnesses in the radial and circumferential directions (16). Based on these results, it has been proposed that human compact bone is orthotropic. The difference between the stiffnesses in the radial and circumferential direction is, however, small and has been attributed to the gradient in porosity going from the periostium to the endostium. Thus, for modeling purposes transverse isotropy is the appropriate choice of symmetry (37,38).

Note that the elastic moduli found at high frequencies by ultrasonic methods are greater than those obtained statically or at low frequencies via mechanical testing machines. This is a result of the rate dependence (viscoelasticity) of bone. In the mechanics of whole bones, the principal macroscopic manifestation of cortical bone tissue anisotropy is that the bending rigidity of a bone (the femur) is much greater than its torsion rigidity (39). The degree of anisotropy and the symmetry of bone tissue depends on the species and on the location of the bone in the body. Bovine plexiform bone is stiffer than human Haversian bone, but dry bone is stiffer than the same type of bone when wet. Bovine plexiform bone is orthotropic and has significantly different elastic moduli in the longitudinal, radial, and circumferential directions (40). These properties reflect the laminar architecture of this type of bone, as shown in cross-section in Fig. 2. The scale mark is in the radial direction and is perpendicular to the laminae. Bovine plexiform bone is a type of primary bone that occurs in relativley young cattle. It is often used for experiments as a result of its availability. Canine femoral bone is also orthotropic (16), however, canine mandibular bone and possibly also human mandibular bone, is transversely isotropic (41).

Anisotropic properties of bone may also be expressed in terms of the technical elastic constants, Young's modulus E, shear modulus G, and Poisson's ratio, v, for different directions. Poisson's ratio is minus the transverse strain divided by the axial strain in the direction of stretching or of compressing force. Representative values for dry human femoral bone are shown in Table 3. The 3 direction is the long axis of the bone, the 2 direction is circumferential, and the 1 direction is radial. We note that Poisson's ratio in isotropic materials must be less than one-half (7). In anisotropic materials, larger values are permissible, so that the values reported for bone do not violate any physical law.

Bone mineral density plays an important role in determining all of the elastic properties described above. Under normal physiological and physical conditions, as bone density increases so do the respective elastic properties. However, bone density alone does not always determine fracture risk in pathologies such as osteporosis. A general term in vogue now, 'bone quality', is being used to qualify what is information is necessary to determine when bone is at risk of failure due to reduced density. Structure–property relationships are the key here, especially in understanding when trabecular bone will fail. Even under the conditions of reduced density, the appropropriate structural organization

of the bone may still provide adequate support during normal function. Similarly, good density alone, if associated with a genetic pathology, such as found in osteopetrosis, will not provide adequate prrotection against fracture. The hardness and elastic moduli of osteopetrotic cortical bone are even well below that of osteoporotic bone even though the density of the former is in the normal range (26,42). Indeed, osteopetrotic bone tends to fracture quite readily.

It is clear that for a detailed modeling of the elastic properties of bone, its complex hierarchical structure must be taken into account (37,38,43).

Strength

The ultimate strength of bone tissue refers to the maximum stress the material can withstand before breaking. The tensile strength of human compact bone (17) in a direction parallel to the osteons is about 150 MPa or 21,000 $\mathrm{lb \cdot in.^{-2}}$ (1 MPa = 145 $\mathrm{lb \cdot in.^{-2}}$). As indicated in Table 5, bone is stronger than various plastics, concrete, brick, some metals, (e.g., aluminum), and most woods. Although bone is stronger than aluminum, commonly used aluminum alloys are stronger than bone. Even so, bone has a lower density than aluminum and a much lower density than steel. The criterion for structural strength in beam bending is material strength divided by the 1.5 power of density [3.×]. For bending of plate-shaped structural elements the criterion is material strength divided by the square of the density. The strength to density ratio for bone is greater than that for structural steel. Therefore bone has very favorable properties in comparison with steel and is competitive with aluminum alloys. Bone is anisotropic in its strength as well as in its elastic behavior. In particular, bone is considerably weaker when loaded transversely than when loaded along the osteon direction. Fortunately, bone in the body does not normally experience significant transverse loads. Several investigators have explored the dependence of bone strength upon age. Tensile strength of adult compact bone decreases 4% per decade of age (44) as does its shear strength (45). In other studies, no significant age dependence of strength was found (46,47). More recently, it was found that the tensile strength of femoral bone decreases 2.1% per decade of age while that of tibial bone decreases 1.2% per decade of age (48). The decrease in strength was statistically significant in the case of femoral bone, but not in the case of tibial bone. Both kinds of bone exhibit significant decreases, 6.8% per decade for femur and 8.4% per decade for tibia, in energy absorption to fracture, a measure of toughness. No significant difference between age-matched males and females was found for any mechanical property (48). Tibial bone is stronger in tension than femoral bone by $\sim$ 19% (49,50) and is $\sim$ 4% stronger than fibular bone.

Currey observes that the average stiffness of human and sheep bone increases monotonically with age (in humans from age 2–50), while energy absorption to fracture decreases. The lower mineralization and stiffness and higher toughness seen in young bones is considered adaptive in view of the many falls and bumps experienced by children and young animals (1).

The fracture behavior of a material is not described fully by its yield and ultimate strengths alone; toughness is also important. Toughness is seen as an important characteristic of bone in view of the microcracks that occur in living bone. Fracture of laboratory specimens containing controlled notches provides information concerning the toughness of materials (49,50). For example, the nominal fracture strength of a tensile specimen with an edge notch of length a is $\sigma_{\mathrm{ult}} = K_{1c} a^{-1/2}/Y$ in which Y depends on the specimen geometry and K_{1c} is called the critical stress intensity factor. A large value of K_{1c} results in high strength even in the presence of a notch; K_{1c} is a measure of material toughness.

The critical stress intensity factor K_{1c} for fracture of compact tension specimens of bovine femur bone is from 2.2 to 4.5 $\mathrm{MN \cdot m^{-3/2}}$, and the specific surface energy for fracture is from 390 to 560 $\mathrm{J \cdot m^{-2}}$ (49). The toughness does not, however depend on the sharpness of the controlled notch in the expected way; the significance of this observation is discussed in the section Compact Bone as a Composite Material. As for the age dependence of bone fracture toughness, the energy absorbed to fracture is observed to decrease during childhood and early middle age, and the elastic modulus increases (51). Recently, fracture surfaces from accident victims have been examined microscopically (52). Such surfaces are observed to be much rougher and more intricate than fracture surfaces produced in laboratory specimens of dead bone, which suggests a greater toughness in living bone. The influence of bone viability on its mechanical properties is not well understood. The dependence of bone elastic modulus on viability has been explored in several studies, but the evidence for a difference in elasticity is not compelling.

Currey (1), compares density, modulus, strength and toughness of bones from deer antler, cow thigh, and whale tympanic bulla. As one might expect, modulus increases and toughness decreases with density. Strength is the highest for the femoral bone. Properties vary considerably between these types of bone; the difference is attributed largely to mineralization but partly to histology. The high mineralization of the whale ear bone is responsible for its stiffness, density and brittleness. The stiffness and density are useful given the acoustic function of the ear bone. The brittleness is not a problem since the ear bone is deep within the skull. Conversely, antler is subjected to repeated severe impacts during use, so toughness is beneficial.

Yielding and Plastic Deformation

Wet bone when loaded sufficiently exhibits a yield point, $\sigma_y = 114$ MPa (13). Beyond this critical stress level, the material does not recover upon the release of the load; permanent or plastic deformation has occurred. In bone as in most materials that exhibit yielding, the yield point is approximated by the proportional limit. The proportional limit is the boundary between the linear and nonlinear portions of the stress–strain curve. The mechanism for yield in bone differs from that in metals. In bone, yield occurs as a result of microcracking and other microdamage while in metals, yield results from motion of dislocations. Published reports have not been in agreement as to the amount of plastic deformation in bone. Much of the disagreement has been attributed to the fact that the observed

plastic deformation is highly sensitive to the hydration of the bone. Dry specimens or specimens with dried surfaces or dried and rewetted surfaces behave in a much more brittle manner than those which have been kept fully hydrated during preparation and testing. In the latter case (9), the maximum strain at fracture of bone under tension in the longitudinal direction is 0.031. Yielding of bone has considerable relevance to the function of bone in the body since very much mechanical energy can be absorbed in plastic deformation without fracture. In certain injuries, plasticity in bone can result in residual deformation, which is obvious on a clinical radiograph (53).

Bone Strain *In Vivo*

To ascertain the significance of bone elasticity and strength data in connection with the function of bone in the body, it is desirable to know what levels of stress and strain occur in bones during normal activities and under traumatic conditions. It is possible to make inferences from macroscopic measurements of forces acting on the extremities. The validity of such inferences is rendered uncertain by the fact that muscle forces cannot generally be uniquely determined. It has become possible to determine bone strains explicitly in various animals (54) and in humans (55) by directly cementing foil strain gages to bone surfaces (56). In a human volunteer, maximum strain along the tibia axis was $\sim 3.5 \times 10^{-4}$ during normal walking at $1.4\ \mathrm{m \cdot s^{-1}}$ and 8×10^{-4} during running at $2.2\ \mathrm{m \cdot s^{-1}}$. Strains of similar magnitudes have been observed in animals such as sheep (53). The largest strain magnitude observed in the normal activity of an animal was 3.2×10^{-3} in the tibia of a galloping horse (57). In comparison (5), in tension in the longitudinal direction human bone yields at a strain of 6.7×10^{-3} and fractures at a strain of 0.03. The strain levels observed *in vivo* are significant in view of the fatigue properties of bone. Race horses can experience bone strain exceeding 4.8×10^{-3} at maximum effort (58).

Fatigue

Bone, like other materials, accumulates damage when loaded repeatedly and this damage can lead to fracture at a lower stress than the ultimate strength measured using a single load cycle. The results of *in vitro* mechanical fatigue studies on dead bone suggest that bone may accumulate significant fatigue damage during normal daily activity. Biological bone remodelling is concluded to be essential to the long-term structural integrity of the skeletal system (59). It is notable that dead bone, unlike steel, does not exhibit an endurance limit. The endurance limit is defined as a stress below which the material can withstand an unlimited number of cycles of repetitive load without breaking in fatigue. In the absence of an endurance limit, dead bone subjected to the prolonged cyclic load of daily activity must eventually break. Living bone, by contrast, is able to repair microdamage generated during fatigue.

The resistance of human cortical bone to fatigue fracture is more strongly controlled by strain range than stress range. Fatigue strength shows a weak positive correlation with bone density and with bone modulus and a weak negative correlation with porosity (59). Fatigue strength in uniaxial tests is lower than in bending tests. In immature bone (60), bone fatigue resistance for a given strain range decreases with maturation, while the elastic modulus, density, and ash content increase with maturation. Cracking or fracture of bone due to fatigue manifests itself clinically as 'stress fractures' that can occur in individuals who suddenly increase their level of physical activity (61). In the case of a person in poor physical condition who sustains a fatigue fracture during military training, it is called a "march fracture". Laboratory results for fatigue in bone are not inconsistent with clinical experience with fatigue fractures in athletes and military recruits (62). In particular, a recruit may accumulate in 6 weeks of training a load history equivalent to 100–1000 miles of very rigorous exercise, associated with peak bone strains of 0.002 and a maximum strain range of 0.004, which can be sufficient to precipitate a fatigue fracture (62).

Stress Concentrations

It is common practice in orthopedic surgery to drill holes in bones for the placement of screws. Such holes cause the bone to be weaker than an intact bone, so that it may fracture as a result of less trauma than would ordinarily be required (63,64). This phenomenon may be understood in view of the fact that holes in elastic solids are known to cause a concentration of stress in the vicinity of the hole. According to the theory of elasticity, the stress concentration is not dependent on a decrease in the amount of material available to bear the load; it can be severe even for small holes. Laboratory studies of whole bone fracture have disclosed that a 3 mm diameter hole weakens a tibia by 40% in bending and 12% in torsion (65). A 2.8-or a 3.6-mm hole reduces the strength of dog bones (63) under rapidly applied torsion by a factor of 1.6. It is of interest to compare the observed stress concentration with predicted values based on the theory of elasticity. The predicted stress concentration factor for a hole in a field of shearing stress is 4.0 provided that the hole is small compared with the structure as a whole (66). The discrepancy has been attributed to the fact that intact bone already has stress raisers by virtue of its heterogeneous structure and porosity (63). However, in specimens loaded in the linear elastic range, well below yield or fracture, distributions and concentrations of strain differ from predicted values (67). Similar discrepancies have been reported in manmade fibrous composites without preexisting porosity (68,69). The role of the fibrous architecture of bone in relation to stress concentrations is discussed in the section on composite properties of bone.

Viscoelasticity

Bone exhibits viscoelastic behavior, that is, the stress depends not only on the strain, but also on the time history of the strain. Such behavior can manifest itself as creep, which is a gradual increase in strain under constant stress; stress relaxation, which is a gradual decrease in stress in a specimen held at constant strain; load-rate dependence of the stiffness; attenuation of sonic or ultrasonic waves; or energy dissipation in bone loaded dynamically (10).

Experimental modalities based on each of the above phenomena have been used in the study of bone (70–74). The results have been converted to a common representation via the interrelationships inherent in the linear theory of viscoelasticity, to permit a direct comparison of results (75,76). In the case of tension–compression, there is very significant disagreement among the published results. This disagreement may result from nonlinear viscoelastic behavior not accounted for in the transformation process, or from experimental artifacts. In the case of shear deformation, however, there is good agreement between results obtained in different kinds of experiments. The loss tangent, which is proportional to the ratio of energy dissipated to energy stored in a cycle of deformation, achieves a minimum value of ~ 0.01 at frequencies from 1 to 100 Hz. At lower and higher frequencies, the loss tangent, hence the magnitude of viscoelastic effects is greater (e.g., 0.08 at 1 MHz and at 1 μHz). To compare, the loss tangent of quartz may be $< 10^{-6}$, in metals, from 10^{-4} to 0.01, in hard plastics from 0.01 to 0.1, and in soft polymers, it may attain values > 1. It is notable that the minimum energy dissipation in bone occurs in a frequency range characteristic of load histories during normal activities.

A synopsis of wet bone viscoelastic behavior in shear is presented in Fig. 7. The tan δ attains a broad minimum over the frequency range associated with most bodily activities. Some authors have suggested that the viscoelastic behavior in bone confers a shock- absorbing role. The observed minimum in damping at frequencies of normal activities is not supportive of such an interpretation.

Some authors refer to three element spring dashpot models used in tutorials on viscoelasticity. The behavior of such a model corresponds to a Debye peak in tan δ, also shown in the figure. This corresponds, by Fourier transformation, to a single exponential in the creep or relaxation behavior. The tan δ of bone occupies a much larger region of the frequency domain than a Debye peak, so the spring dashpot model is not appropriate.

Compact Bone as a Composite Material

Early composite models (83–87) for bone were two-phase models involving the mineral and protein phases only. At the ultrastructural level one may imagine the mineral crystals as a particulate reinforcing phase and the surrounding collagen as a matrix phase. Strong arguments were presented that bone cannot be described simply as a compound bar or as a material similar to prestressed concrete. Based on measured Young's moduli of 114 GPa for hydroxyapatite (24) and 1.2 GPa for collagen derived from tendon (83), rigorous upper and lower bounds on the elastic modulus of compact bone were calculated (85,86). Young's modulus of bone lay between these bounds and was less than the value obtained by a simple rule of mixtures approach. Similar comparisons were made for other natural collagen-apatite composites such as dentin and enamel from human teeth.

The wide disparity in the upper and lower bounds in this approach limited its applicability. Later, a composite model of the elastic properties of cortical bone was attempted in order to explain the angular dependence of the elastic properties of both wet and dried bovine cortical bone as determined in an ultrasonic wave propagation experiment (87). However, this modeling failed due to the much stronger decay in the angular dependence of the elastic modulus calculated than was found experimentally. Both this

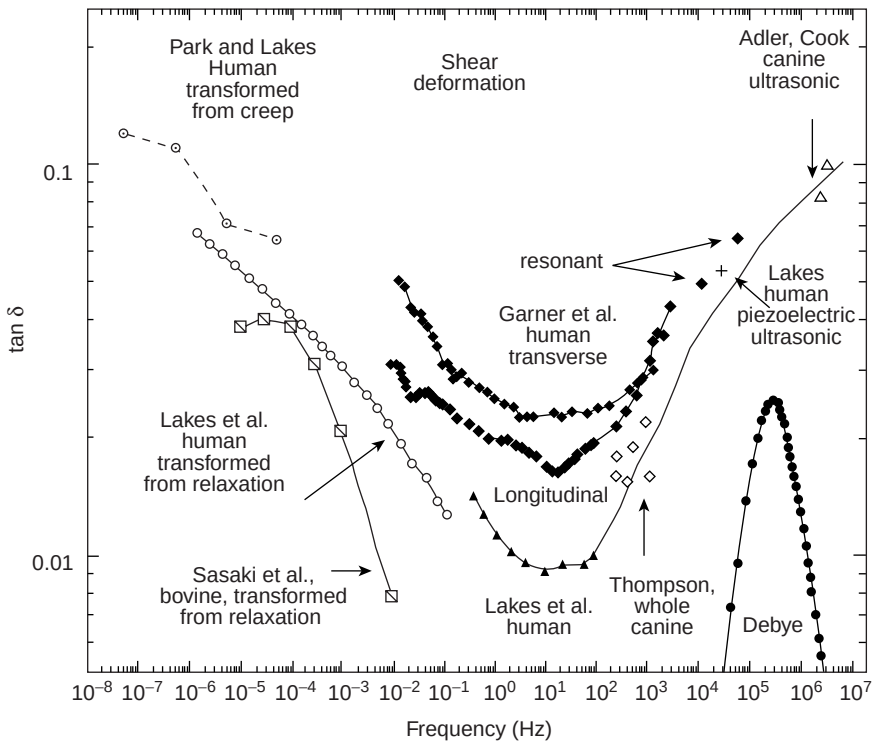


Figure 7. Viscoelastic behavior of wet compact bone in shear. Comparison of results of various authors, adapted from (3). Low frequency tan δ inferred from slope of long-term creep by Park and Lakes, 1986 (77), center circles, ⊙. Results calculated from integration of constitutive equation of Sasaki et al. 1993 (78) for torsion relaxation in bovine bone (slant squares, □). Damping adapted from data of Lakes et al., 1979 (75) for wet human tibial bone at 37°C (Calculated from relaxation, circles, ○; directly measured, ▲). Direct tan δ measurements by Garner et al., 2000 (79) for wet human bone in torsion, diamonds, →. Damping data of Thompson, 1971 (80) for whole dog radius at acoustic frequencies (diamonds, □). Damping of wet human femoral bone by Lakes, 1982 (81) via a piezoelectric ultrasonic oscillator (cross, +). Damping at ultrasonic frequency for canine bone by Adler and Cook (1975) (82) at room temperature (open triangles, △). Theoretical debye peak corresponding to an exponential in the time domain, solid circles •.

attempt (88) and the earlier calculations (84–87) failed to model the properties of compact bone because they did not incorporate the dependence of its properties on its complex hierarchical structure. Inclusion of the hierarchical structural organization of bone was accomplished by adaptation of the hollow fiber composite model (89) so that it resembled the structure of Haversian bone with the osteons viewed as hollow fibers embedded in a matrix (36,37). Subsequently, the same model (35,36,89) was adapted to calculate the viscoelastic properties of bone (90). In recent years, homogenization techniques have been applied to obtain even further improvements In the composite modeling (91,92).

The matrix is the ground substance that comprises the cement lines between osteons and is thought to be principally composed of mucopolysaccharides. The stiffness of the ground substance has not been determined experimentally, but it has been inferred from the composite model in conjunction with experimental data from the ultrasound studies on whole bone (93,94). Based on this calculation, the ground substance was computed to be about one-quarter as stiff as the osteon itself. The view of the ground substance as a compliant interface is also supported by the results of several experimental studies. Localized slippage occurs at the cement lines in specimens of bovine plexiform bone (95) of human Haversian bone (77) subjected to prolonged stress. Under such conditions, the ground substance at the cement lines appears to behave in a viscous manner. Considering the full range of load rates and frequencies, one may view the ground substance as a compliant and highly viscoelastic material. Such a conclusion is perhaps surprising (2) in view of the fact that the ground substance is highly mineralized (96). Nevertheless a view of the ground substance as a compliant interface is supported by further experiments. For example, studies of single osteons and osteon groups in torsion have revealed size effects to occur and the osteon to have a higher effective shear modulus than whole bone (97). Similar size effects were observed in torsional and bending experiments upon larger microsamples (98,99). These latter results have been interpreted in light of a generalized form of elasticity theory, known as Cosserat elasticity, which admits both strain of the material and local rotations of microscopic constituents, for example, the osteons (97–99). A twist per unit area or couple stress can occur in addition to a force per unit area or couple stress. By contrast, classical elasticity (Eq. 1), which describes most ordinary materials, involves only strains and stresses. Cosserat elasticity is likely differ significantly from classical elasticity in its predictions of stress and strain around holes, cracks, and interfaces. For large, whole bones, conventional anisotropic elasticity has been shown to be entirely adequate (39).

The interface between osteons also appears to be important in conferring a measure of fracture toughness upon compact bone. In particular, the crack blunting mechanism of Cook and Gordon (100) confers toughness in fibrous media by the action of a *weak* interface between fibers in blunting a propagating crack (101). Evidence for the role of the cement substance as such a weak interface has been presented by Piekarski (102). Pullout of fibers can result in a large energy absorption in the fracture of fibrous composites (103). This toughening mechanism also appears to be operative in compact bone, as seen in micrographs of pullout of osteons in fractured bone specimens (51,104).

Consideration of bone as a composite material may provide insight regarding various phenomena in the mechanics of bone. In particular, stress concentration around holes is significantly less than the predicted value (63). The strength of notched specimens does not decrease with the sharpness of the notch as expected; instead, the strength is independent of notch sharpness (49). Residual strain around holes in bone does not follow the predictions of classical anisotropic elasticity (98,99). The fatigue life of bone specimens in bending exceeds that of specimens in tension by a factor of several thousand (59). Similar effects have been observed in various manmade fibrous composites such as boron-epoxy and graphite-epoxy (104). Such phenomena are not correctly predicted by elasticity theory, but may be accounted for in the context of structural models, composite theories, or generalized continuum models (77).

Polycrystalline elastic properties and single-crystal elastic constants of apatites that are useful In the modeling of calcified tissues elasticity are given in Table 6.

Mechanical Properties of Cancellous Bone

Cancellous or trabecular bone is a highly porous or cellular form of bone. In a typical long bone, the cortex or exterior of the shaft (diaphysis) and flared ends (metaphysis) is composed of compact bone while the interior, particularly near the articulating ends, is filled with cancellous bone. Cancellous bone may also be found to fill the interior of short bones and flat bones as well as in the interior of bony tuberosities under mucle attachments. The structure of cancellous bone is that of a latticework of bars and plates; typical structure is shown in Fig. 3. The volume fraction of solid material can be from 5 to 70%; the interstices are filled with marrow. The compressive strength σ_{ult}(in MPa) depends very much on the density ρ (in $g \cdot cm^{-3}$) and also varies with the strain rate de/dt (in s^{-1}) as follows (105).

$$\sigma_{ult} = 68(de/dt)^{0.06}\rho^2$$

This relation also models the compressive strength of compact bone (221 MPa, at a density of $\rho. = 1.8\ g \cdot cm^{-3}$). To a certain degree of approximation, both compact and cancellous bone may be mechanically viewed as a single material of variable density (105). Density is not, however, the only determinant of the properties of cancellous bone. The microstructure can vary considerably from one part of the body to another (106). For example, in the vertebras and in the tibia (107), a highly oriented, columnar architecture is observed. This kind of trabecular bone is highly anisotropic: the Young's modulus in the longitudinal direction can exceed that in the transverse direction by more than a factor of 10 (107). By contrast, in regions such as the proximal part of the bovine humerus, the cancellous bone can be essentially isotropic (108). This bone is about twice as strong in compression as in tension. In many ways, cancellous bone (109–111) is similar in its behavior to manmade rigid cellular foams (112). For example, in compression, the stress–strain curves contain a linear elastic region, up to a strain of $\sim$0.05, at which the cell walls bend or compress (112). A plateau

region of almost constant macroscopic stress is associated with elastic buckling, plastic yield, or fracture of the cell walls. The compressive failure of the cancellous bone proceeds at approximately constant stress until the cell walls touch each other; at this point any further compression causes the stress to rise rapidly (112). By contrast, fracture of cancellous bone in tension proceeds abruptly and catastrophically (108). The energy absorption capacity of cancellous bone is consequently much less in tension than it is in compression (109). This suggests that tensile and avulsion fractures of cancellous bone observed clinically are associated with minimal energy absorption, and therefore may be precipitated by relatively minor trauma (109). The elastic modulus E for cancellous bone increases as the square of the density ($E = k\rho^2$) if the structure consists of open cells forming a network of rods (110,111). In closed-cell cancellous structures consisting of plates, the modulus E is proportional to the cube of the density ($E = k\rho^3$). Based on study of micrographs and density maps of femora and vertebras, it is suggested that an open cell structure of rods is found when the solid volume fraction is less than ~ 0.13, while a closed cell plate structure occurs at a density of $>$ 350 $kg \cdot m^{-3}$ corresponding to a relative density or solid volume fraction of 0.20 (110). The relationship between density and structure may not be so straightforward in all situations. As for the highly oriented columnar cancellous bone from the human tibia (107), the modulus E in the longitudinal direction is proportional to the density ($E = k\rho$). Such behavior is anticipated on the basis of an axial compressional mode of deformation of the cell walls, in contrast to the bending mode that is expected in rod and plate structures (112).

It is important to distinguish the difference between the elastic moduli of volumetric samples of trabecular bone that are highly porous, low density structures, thus exhibiting low moduli, generally well below 1.0 GPa as obtained by mecanical testing techniques, and that of individual trabeculae, comparable in structure and density to cortical bone,thus exhibiting much higher moduli (10–20 GPa), as obtained in recent years by nanoindentation (19) and SAM (18) in addition to that obtained by mechanical testing). Values of the elastic moduli of both trabecular volumes and individual trabeculae are given in Table 7.

ADAPTIVE PROPERTIES OF BONE

Phenomenology

The relationship between the mass and form of a bone to the forces applied to it was appreciated by Galileo (113), who is credited with being the first to understand the balance of forces in beam bending and with applying this understanding to the mechanical analysis of bone. Wolff (114) published his seminal 1892 monograph on bone remodeling; the observation that bone is reshaped in response to the forces acting on it is presently referred to as Wolff's law. Cowin, in Chapter 25 of Ref. 3, discussed "The Problems with Wolff's Law ". Many relevant observations regarding the phenomenology of bone remodeling have been compiled and analyzed by Frost (115,116). Salient points are as follows:

1. Remodeling is triggered not by principal stress but by "flexure".
2. Repetitive dynamic loads on bone trigger remodelling; static loads do not.
3. Dynamic flexure causes all affected bone surfaces to drift toward the concavity that arises during the act of dynamic flexure.

These rules are essentially qualitative and they do not deal with underlying causes. A critique of these ideas has been presented by Currey (1,2). Additional aspects of bone remodeling may be found in the clinical literature. For example, after complete removal of a metacarpal and its replacement with graft consisting of a strut of tibial bone, the graft becomes remodeled to resemble a real metacarpal; the graft continues to function after 52 years (117). In the standards of the Swiss Association for Internal Fixation it is pointed out that severe osteoporosis can result from the use of two bone plates in the same region as a result of the greatly reduced stress in the bone (118). Pauwels (119) suggested that as a result of bending stresses the medial and lateral aspects of the femur should be stiffer and stronger than the anterior and posterior aspects. Such a difference has actually been observed (120). Large cyclic stress causes more resorption than large static stress (121). Immobilization of humans causes loss of bone and excretion of calcium and phosphorus (122). Long spaceflights under zero gravity also cause loss of bone (123,124); hypergravity induced by centrifugation strengthens the bones of rats (125,126). Studies of stress-induced remodeling of living bone have been performed *in vitro* (127). Recently, *in vivo* studies in pigs (128) were conducted. In this study, strains were directly measured by strain gages before and after remodeling. Remodeling was induced by removing part of the pigs' ulna so that the radius bore all the load. Initially, the peak strain in the ulna approximately doubled. New bone was added until, after 3 months, the peak strain was about the same as on the normal leg bones. *In vivo* experiments conducted in sheep (129) have disclosed similar results. It is of interest to compare the response time noted in the above experiments with the rate of bone turnover in healthy humans. The life expectancy of an individual osteon in a normal 45 year old man is 15 years and it will have taken 100 days to produce it (130,131).

Remodeling of Haversian bone seems to influence the quantity of bone but not its quality, that is, young's modulus, tensile strength, and composition (132). However, the initial remodeling of primary bone to produce Haversian bone results in a reduction in strength (1,2). As for the influence of the rate of loading on bone remodeling, there is good evidence to suggest that intermittent deformation can produce a marked adaptive response in bone, whereas static deformation has little effect (127). Experiments (133) upon rabbit tibiae bear this out. In the dental field, by contrast, it is accepted that static forces of long duration move teeth in the jawbone. In this connection (134), the direction (as well as the type) of stresses acting on the bone tissue should also be considered. Currey (1,2) points out that the response of different bones in the same skeleton to mechanical loads must differ, otherwise lightly loaded

bones such as the top of the human skull, or the auditory ossicles, would be resorbed.

Failure of bone remodeling to occur normally in certain disease states is of interest: for example, osteropetrotic bone contains few if any viable osteocytes and usually contains a much larger number of microscopic cracks than adjacent living bone (135). This suggests that the osteocytes play a role in detecting and repairing the damage. In senile osteoporosis, bone tissue is removed by the body, often to such an extent that fractures occur during normal activities. Osteoporosis may be referred to as a remodeling error (116).

Some theoretical work, notably by Cowin and others (3,136) has dealt with the problem of formulating Wolff's law in a quantitative fashion. In this theory, constitutive equations are developed, which predict the remodeling response to a given stress. Stability considerations are invoked to obtain some constraints on the parameters in the constitutive equation.

Feedback Mechanisms

Bone remodeling appears to be governed by a feedback system in which the bone cells sense the state of strain in the bone matrix around them and either add or remove bone as needed to maintain the strain within normal limits. The process or processes by which the cells are able to sense the strain and the important aspects of the strain field are presently unknown. Bassett and Becker (137) reported that bone is piezoelectric, that is, that it generates electric fields in response to mechanical stress; they advanced the hypothesis that the piezoelectric effect is the part of the feedback loop by which the cells sense the strain field. This hypothesis obtained support from observations of osteogenesis in response to externally applied electric fields of the same order of magnitude as those generated naturally by stress via the piezoelectric effect. The study of bone bioelectricity has received impetus from observations that externally applied electric or electromagnetic fields stimulate bone growth (138). The electrical hypothesis, while favored by many, has not been proven. Indeed, other investigators have advanced competing hypotheses that involve other mechanisms by which the cells are informed of the state of stress around them.

For example, inhomogeneous deformation at the lamellae may impinge on osteocyte processes and thus trigger the osteocytes to initiate bone formation or remodelling (139). Motion at the cement lines was observed and it was suggested that such motion could act as a passive mechanism by which bone's symmetry axes may become aligned to the direction of time averaged principal stresses (99). Stress on bone may induce flow of fluid in channels, (e.g., canaliculi), and such flow could play a role in the nutrition and waste elimination of osteocytes, which may be significant in bone remodeling (140). In a related vein, theoretical arguments have been presented in support of the hypothesis that bone cells are directly sensitive to hydrostatic pressure transmitted to them from the bone matrix via the tissue fluid (141). Although no experimental test of this direct pressure hypothesis has been published, we observe with interest that direct hydrostatic pressure has been observed to alter the swimming behavior of paramecia, possibly by means of action upon the cell membrane (142). Otter and Salman found that a hydrostatic pressure of 68 atm abolishes the reversing of direction of swimming, 170 atm stops swimming, and 400–500 atm irreversibly damages the cell. We observe that 100 atm corresponds to 1400 psi stress, or in bone, a strain of 0.07%, which is in the normal range of bone strain and 500 atm corresponds to 7000 psi or a strain of 0.35%, well above the normal range of bone strain. Stress in bone also results in temperature differences between osteons (143); the cells may be sensitive to sudden temperature changes during human activity. A mechanochemical hypothesis has been advanced, in which the solubility of calcium may be affected by stress in the bone matrix (144). Strain energy in bone might also influence the energetics of bone mineral nucleation (145). It has also been suggested that remodeling may be initiated in response to microcracks generated by mechanical fatigue of bone (146). In summary, many hypotheses have been proposed for the mechanism by which appropriate cells sense the state of strain in bone, but little or no experimental evidence is available to discriminate among them.

Cellular and Biochemical Aspects of Bone Remodeling

The adaptive response of bone to mechanical stimuli is mediated by living cells. A great deal is known concerning bone cell function and its control by ionic and hormonal factors, but little is known concerning the effect of mechanical strain in bone upon the biochemistry of its cells. Rasmussen and Bordier (147) have presented an extensive review of studies of bone cell physiology. Recently, the biochemical consequences of electrical stimulation of bone have been reported (148). Biochemical steps associated with cell activation are as yet poorly understood, but ion fluxes appear to play a role (149). Cyclic nucleotides mediate the effects of extracellular signals (150) and prostaglandins modulate them (149). Prostaglandin E_2 has been hypothesized to mediate bone resorption in trauma, malignancy, and periodontal disease. This prostaglandin, as well as the cellular constituents cyclic AMP and cyclic GMP, has been found in association with regions of bone stimulated electrically (148).

ELECTRICAL PROPERTIES OF BONE

Fukada and Yasuda (151) first demonstrated that dry bone is piezoelectric in the classic sense, that is, mechanical stress results in electric polarization, the indirect effect; and an appplied electric field causes strain, the converse effect. The piezoelectric properties of bone are of interest in view of their hypothesized role in bone remodeling (137). Wet collagen, however, does not exhibit piezoelectric response. Studies of the dielectric and piezoelectric properties of fully hydrated bone raise some doubt as to whether wet bone is piezoelectric at all at physiological frequencies (152). Piezoelectric effects occur in the kilohertz range, well above the range of physiologically significant frequencies (152). Both the dielectric properties (153) and the piezoelectric properties of bone (154) depend strongly on

frequency. The magnitude of the piezoelectric sensitivity coefficients of bone depends on frequency, on direction of load, and on relative humidity. Values up to 0.7 pC/N have been observed (154), to be compared with 0.7 and 2.3 pC/N for different directions in quartz, and 600 pC/N in some piezoelectric ceramics. It is, however, uncertain whether bone is piezoelectric in the classic sense at the relatively low frequencies which dominate in the normal loading of bone. The streaming potentials examined originally by Anderson and Eriksson (155,156) can result in stress generated potentials at relatively low frequencies even in the presence of dielectric relaxation, but this process is as yet poorly understood.

Potentials observed in bent bone differ from predictions based on the results of experiments performed in compression (157). The piezolectric polarization may consequently depend on the strain gradient (157) as well as on the strain. This piezoelectric theory has been criticized as ad hoc by some authors, however, the idea has some appeal in view of Frost's modeling (115,116) and Currey's suggestion (1,2) that strain gradients may be significant in this regard. The gradient theory is not ad hoc, but can be obtained theoretically from general nonlocality considerations (158). The physical mechanism for such effects is hypothesized to lie in the fibrous architecture of bone (26,78). Theoretical analyses of bone piezoelectricity (159–162) may be relevant to the issue of bone remodeling. Recent thorough studies have explored electromechanical effects in wet and dry bone. They suggest that two different mechanisms are responsible for these effects: Classical piezoelectricity due to the molecular asymmetry of collagen in dry bone, and fluid flow effects, possibly streaming potentials in wet bone (163).

Bone exhibits additional electrical properties which are of interest. For example, the dielectric behavior (e.g., the dynamic complex permittivity) governs the relationship between the applied electric field and the resulting electric polarization and current. Dielectric permittivity of bone has been found to increase dramatically with increasing humidity and decreasing frequency (152,153). For bone under partial hydration conditions, the dielectric permittivity (which determines the capacitance) can exceed 1000 and the dielectric loss tangent (which determines the ratio of conductivity to capacitance) can exceed unity. Both the permittivity and the loss are greater if the electric field is aligned parallel to the bone axis. Bone under conditions of full hydration in saline behaves differently: the behavior of bovine femoral bone is essentially resistive, with very little relaxation (164). The resistivity is $\sim$ 45–48 μm for the longitudinal direction, and three to four times greater in the radial direction. These values are to be compared with a resistivity of 0.72 μm for physiological saline alone. Since the resistivity of fully hydrated bone is $\sim$ 100 times greater than that of bone under 98% relative humidity, it is suggested that at 98% humidity the larger pores are not fully filled with fluid (164).

Compact bone also exhibits a permanent electric polarization as well as pyroelectricity, which is a change of polarization with temperature (165,166). These phenomena are attributed to the polar structure of the collagen molecule; these molecules are oriented in bone. The orientation of permanent polarization has been mapped in various bones and has been correlated with developmental events.

Electrical properties of bone are relevant not only as a hypothesized feedback mechanism for bone remodeling, but also in the context of external electrical stimulation of bone to aid its healing and repair (167,168).

The frequency of electrical stimulation influences its effectiveness (169). A frequency band of 20–30 Hz was found from analysis of strain data. Bone growth could be stimulated more easily in the avian ulna at relatively higher frequencies. Electromagnetic stimuli also prevent bone loss due to disuse as revealed in an isolated canine fibula model (170). Bone density may be maintained and increased by weight lifting, which involves no medical intervention. Indeed, bone mineral content values (as determined with dual photon absorptiometry) in the spines of athletes were extremely high and were closely correlated to the amount of weight lifted during training (171).

BIBLIOGRAPHY

Cited References

1. Currey J. The mechanical adaptations of bones. Princeton: Princeton University Press; 1984.
2. Currey J. Bone Structure and Mechanics. Princeton: Princeton University Press; 2002.
3. Cowin S. Bone Mechanics. 2nd ed. Boca Raton, FL: CRC Press; 2001.
4. Park JB, Lakes RS. Biomaterials. 2nd ed. New York: Plenum; 1992.
5. Hancox NM. Biology of Bone. Cambridge, (MA): Cambridge University Press; 1972.
6. LeGeros RZ. Monographs in Oral Science. Karger: 1991.
7. Sokolnikoff IS. Mathematical theory of elasticity. Krieger; 1983.
8. Ferracane JL. Materials In Dentistry. 2nd ed. Philadelphia: Lippincott, Williams & Wilkins; 2001.
9. Briggs A. Acoustic Microscopy. Oxford: Clarendon Press; 1992.
10. Lakes RS. Viscoelastic Solids. Boca Raton, FL: CRC Press; 1998.
11. Eppell SJ, Tong WL, Katz JL, Kuhn L, Glimcher MJ. Shape and size of isolated bone mineralites measured using atomic force microscopy" J Ortho Res 2001;19:1027–1034.
12. Tong W, Glimcher MJ, Katz JL, Kuhn L, Eppell SJ. Size and shape of Mineralites in young bovine bone measured by atomic force microscopy. Calcif Tiss Inter 2003;72:592–598.
13. Reilly DT, Burstein AH. The elastic and ultimate properties of compact bone tissue. J Biomech 1975;8:393–405.
14. Craig RG, Peyton FA. Elastic and mechanical properties of human dentin. J Dental Res 1958;37:710–718.
15. Craig RG, Peyton FA. Compressive properties of enamel, dental cements, and gold. J Dental Res 1961;40:936–945.
16. Ashman RB, Cowin SC, Van Buskirk WC, Rice JC. A continuous wave technique for the measurement of the elastic properties of cortical bone. J Biomech 17:349–361.
17. Reilly DT, Burstein AH. The mechanical properties of cortical bone. J Bone Jnt Surg 1974;56A:1001–1022.
18. Bumrerraj S, Katz JL. Scanning acoustic microscopy study of human cortical and trabecular bone. Ann Biomed Eng 2001;29:1–9.
19. Rho JY, Pharr GM. Effects of drying on the mechanical poroperties of bovine femur measured by nanoindentation. J Mater Sci, Matyer Med 1999;10:1–4.

20. Kapur R. The use of scanning acoustic microscopy to study the microstructural properties of the dentin/enamel junction, B.S. Project (Katz JL, Advisor), Department of Biomedical Engineering, Case Western Reserve University; 1999.
21. Löst C, Irion KM, Nussle JC. Two-dimensional distribution of sound velocity in ground sections of enamel. Endod Dent Traumatol 1992;8:215–218.
22. Van Buskirk WC, Ashman RB. The elastic moduli of bone. In: Mechanical Properties of Bone, Joint ASME-ASCE Applied Mechanics, Fluids Engineering and Bioengineering Conference. Boulder, CO; 1981.
23. Lees S, Rollins FR, Jr. Anisotropy in hard dental tissues. J Biomech 1972;5:557–566.
24. Lees S, Ahern JM, Leonard M. Parameters influencing the sonic velocity in compact calcified tissues of various species. J Acoust Soc Am 1983;74:28–33.
25. Gilmore RS, Pollack RP, Katz JL. The elastic properties of bovine dentin and enamel. Arch Oral Biol 1970;15:787– 796.
26. Katz JL, Lipson S, Yoon HS, Maharidge R, Meunier A, Christel P. The effects of remodeling on the elastic properties of bone, Calc Tiss Inter 1984;36:S31–S36.
27. Yoon HS, Katz JL. Ultrasonic wave propagation in human cortical bone. II. Measurements of elastic properties and microhardness. J Biomech 1976;9:459–464.
28. Yoon HS, Newnham RE. Elastic Properties of fluorapatite. Am Min 1969;54:1193–1197.
29. Gordon JE. Structures. Penguin; 1983.
30. Lemons RA, Quate CF. Acoustic microscopy-scanning version. Appl Phys Lett 1974;24:163–165.
31. Lemons RA, Quate CF. Acoustic microscopy. Phys Acoust 1979;14:1–92.
32. Katz JL, Meunier A. Scanning acoustic microscope studies of the elastic properties of osteons and osteon lamellae. J Biomech Eng 1993;115:543–548.
33. Evans FG, Bang S. Differences and relationships between the physical properties and the microscopic structure of human femoral, tibial, and fibular bone. Am J Anat 1967;120:79–88.
34. Evans FG, Vincentelli R. Relations of the compressive properties of human cortical bone to histological structure and calcification. J Biomech 1974;7:1–10.
35. Lang SB. Ultrasonic method for measuring elastic coefficients of boneand results on fresh and dried bovine bones. IEEE Trans Biomed Eng, BME 1970;17:101–105.
36. Yoon HS, Katz JL. Ultrasonic wave propagation in human cortical bone, I. Theoretical considerations for hexagonal symmetry, J Biomech 1976;9:407–412.
37. Katz JL. Hierarchical modeling of compact Haversian bone as a fiber reinforced material. In: 1976 Advances in Bioengineering. New York City: ASME; 1976. p 17–18.
38. Katz JL. On the anisotropy of young's modulus of bone. Nature (London) 1980;283:106–107.
39. Huiskes R, Janssen JD, Sloof TJ. A detailed comparison of experimental and theoretical stress analyses of a human femur. In: Mechanical Properties of Bone, Joint ASME-ASCE Applied Mechanics, Fluids Engineering and Bioengineering Conference. Boulder, (CO); 1981.
40. Lipson SF, Katz JL. The relationship between the elastic properties and microstructure of bovine cortical bone. J Biomech 1984;17:231–240.
41. Ashman RB, Rosina G, Cowin SC, Fontenot MG. The bone tissue of the canine mandible is elastically isotropic. J Biomech 1985;18:717–721.
42. Ashman RB, Van Buskirk WC, Cowin SC, Sandbornj PM, Wells MK, Rice JC. The mechanical Properties of immature osteopetrotic bone. Calc Tiss Inter 1985;37:73–76.
43. Lakes RS. Materials with structural hierarchy. Nature (London) 1993;361:511–515.
44. Melick RA, Miller DR. Variations of tensile strength of human cortical bone with age. Clin Sci 1966;30:243–248.
45. Hazama H. Study of the torsional strength of the compact substance of human beings. J Kyoto Pref Med Univ 1956;60:167–184.
46. Evans FG, Lebow M. Regional differences in some of the physical properties of the human femur. J Appl Physiol 1951;3:563–572.
47. Sedlin ED, Hirsch C. Factors affecting the determination of the physical properties of femoral cortical bone. Acta Orthop Scand 1966;37:29–48.
48. Burstein AH, Reilly DT, Martens M. Aging of bone tissue: mechanical properties. J Bone Jnt Surg 1976;58A:82–86.
49. Bonfield W, Datta PK. Fracture toughness of compact bone. J Biomech 1976;9:131–134.
50. Behiri JC, Bonfield W. Crack velocity dependence of longitudinal fracture in bone. J Mat Sci 1980;15:1841–1849.
51. Currey JD. Changes in the impact energy absorption of bone with age. J Biomech 1979;12:459–469.
52. Corondan G, Haworth WL. A fractographic study of human long bone. J Biomech 1986;19:207–218.
53. Carter DR, Spengler DM. Biomechanics of fracture. In: Sumner Smith G, editor. Bone in Cinical Orthopaedics. New York: Saunders; 1982. p 305–332.
54. Lanyon LE. Analysis of surface bone strain in the calcaneus of sheep during normal locomotion. J Biomech 1973;6:41– 69.
55. Lanyon LE, Hampson WGJ, Goodship AE, Shah JS. Bone deformation recorded in vivo from strain gauges attached to the human tibial shaft. Acta Orthop Scand 1975;46:256–268.
56. Caler WE, Carter DR, Harris WH. Techniques for implementing an *in vivo* bone strain gage system. J Biomech 1981;14:503–507.
57. Rubin CT. Skeletal strain and the functional significance of bone architecture. Calcif Tissue Int 1984;36:S11–S18.
58. Nunamaker DM, Butterweck DM, Provost MT. Fatigue fractures in thoroughbred racehorses: relationships with age, peak bone strain, and training. J Orthop Res 1990;8:694.
59. Carter DR, Caler WE, Spengler DM, Frankel VH. Uniaxial fatigue of human cortical bone. The influence of tissue physical characteristics. J Biomech 1981;14:461–470.
60. Keller TS, Lovin JD, Spengler DM, Carter DR. Fatigue of immature baboon cortical bone. J Biomech 1985;18:297–304.
61. Devas MB. Stress fractures. Churchill Livingstone, London; 1975.
62. Carter DR, Caler WE, Spengler DM, Frankel VH. Fatigue behavior of adult cortical bone: the influence of mean strain and strain range. Acta Orthop Scand 1981;52:481–490.
63. Brooks DB, Burstein AH, Frankel VH. The biomechanics of torsional fractures: the stress concentration effect of a drill hole. J Bone Jnt Surg 1970;52A:507–514.
64. Burstein AH, Currey JD, Frankel VH, Heiple KG, Lunseth P, Vessely JC. Bone strength: the effect of screw holes. J Bone Jnt Surg 1972;54A:1143–1156.
65. Laurence M, Freeman MA, Swanson SA. Engineering considerations in the internal fixation of fractures of the tibial shaft. J Bone Jnt Surg 1969;51B:754–768.
66. Timoshenko S, Goodier JM. Theory of elasticity, 3rd ed. New York: McGraw Hill; 1983.
67. Lakes RS, Yang JFC. Concentration of strain around holes in a strip of compact bone, Developments in mechanics, Proceedings of the 18th Midwestern Mechanics Conference. Volume 12, Iowa City; 1983. p 233–237.
68. Awerbuch J, Madhukar S. Notched strength of composite laminates: predictions and experiments, a review. J Reinforced Plast Comp 1985;4:3–159.
69. Daniel IM. Strain and failure analysis in graphite/epoxy plates with cracks. Exper Mech 1978;18:246–252.

70. Smith R, Keiper D. Dynamic measurement of viscoelastic properties of bone. Am J Med Electr 1965;4:156.
71. Currey JD. Anelasticity in bone and echinoderm skeletons. J Exper Biol 1965;43:279.
72. Black J, Korostoff E. Dynamic mechanical properties of viable human cortical bone. J Biomech 1973;16:435.
73. Tennyson RC, Ewert R, Niranjan V. Dynamic viscoelastic response of bone. Exp Mech 1972;12:502.
74. Lugassy AA, Korostoff E. Viscoelastic behavior of bovine femoral cortical bone and sperm whale dentin. In: Research in Dental and Medical Materials. New York: Plenum; 1969.
75. Lakes RS, Katz JL, Sternstein SS. Viscoelastic properties of cortical bone: Part 1: Torsional and biaxial studies. J Biomech 1979;12:657.
76. Lakes RS, Katz JL. Interrelationships among the viscoelastic functions for anisotropic solids: application to calcified tissues and related systems. J Biomech 1974;17:259.
77. Park HC, Lakes RS. Cosserat micromechanics of human bone: strain redistribution by a hydration-sensitive constituent. J Biomech 1986;19:385–397.
78. Sasaki N, Nakayama Y, Yoshikawa M, Enyo A. Stress relaxation function of bone and bone collagen. J Biomech 1993;26:1369–1376.
79. Garner E, Lakes RS, Lee T, Swan C, Brand R. Viscoelastic dissipation in compact bone: implications for stress-induced fluid flow in bone. J Biomech Eng 2000;122:166–172.
80. Thompson G, Experimental studies of lateral and torsional vibration of intact dog radii, [dissertation]. Stanford (CA): Stanford University; 1971.
81. Lakes RS. Dynamical study of couple stress effects in human compact bone. J Biomech Eng 1982;104:6–11.
82. Adler L, Cook CV. Ultrasonic parameters of freshly frozen dog tibia. J Acoust Soc Am 1975;58:1107–1108.
83. Currey JD. Three analogies to explain the mechanical properties of bone. Biorheology 1964;2:1–10.
84. Welch DO. The composite structure of bone and its response to mechanical stress. Recent Adv Eng Sci 1970;5:245–262.
85. Katz JL. Hard tissue as a composite material-I. Bounds on the elastic behavior. J Biomech 1971;4:455–473.
86. Piekarski K. Analysis of bone as a composite material. Inter J Eng Sci 1973;11:557–565.
87. Currey JD. The relationship between the stiffness and the mineral content of bone. J Biomech 1969;2:477.
88. Bonfield W, Grynpas MD. Anisotropy of Young's modulus of bone. Nature (London) 1977;270:453–454.
89. Hashin Z, Rosen BW. The elastic moduli of fiber reinforced materials. J Appl Mech 1964;31:223–2xx.
90. Gottesman T, Hashin Z. Analysis of viscoelastic behavior of bone on the basis of microstructure. J Biomech 1979;13:89–yy.
91. Hogan H. Micromechanics modeling of haversian cortical bone properties. J Biomech 1992;25:549–zzz.
92. Crolet JM, Aoubiza B, Meunier A. Compact bone: Numerical simulation of mechanical characteristics. J Biomech 1993;26:677–aaa.
93. Katz JL, Maharidge RL, Yoon HS. The estimation of interosteonal mechanical properties from a composite model for haversian bone. In: Perren SM, Scheider E, editors. Biomechanics: Current Interdisciplinary Research. Dordrecht: Martinus Injhoff, 1985. p 179–184.
94. Katz JL, Maharidge RL, Yoon HS. Calculation of inter-osteonal mechanical properties for haversian bone based on a hierarchical composite model. Biomechanics Symp. AMD-Vol. 68, FED-Vol. 21. New York City: ASME; 1985. p 33–35.
95. Lakes RS, Saha S. Cement line motion in bone. Science 1979;204:501–503.
96. Frasca P. Scanning-electron microscopy studies of 'ground substance' in the cement lines, resting lines, hypercalcified rings, and reversal lines of human cortical bone, Acta Anatomica 1981;109:115–121.
97. Frasca P, Harper R, Katz JL. Strain and frequency dependence of shear storage modulus for human single osteons and cortical bone microsamples-size and hydration effects. J Biomech 1981;14:679–690.
98. Yang JFC, Lakes RS. Transient study of couple stress effects in human compact bone: Torsion. J Biomech Eng 1981;103:275–279.
99. Yang JFC, Lakes RS. Experimental study of micropolar and couple stress elasticity in compact bone in bending. J Biomech 1982;15:91–98.
100. Cook J, Gordon JE. A mechanism for the control of crack propagation in all-brittle systems. Proc R Soc London 1964;A282:508–520.
101. Kelly A. The strengthening of metals by dispersed particles. Proc R Soc London 1964;A282:63–79.
102. Piekarski K. Fracture of Bone. J Appl Phys 1970;41:215–223.
103. Kelly A. Strong solids. London: Oxford University Press; 1966.
104. Wright TM, Barnett DM, Hayes WC. Residual stresses in bone. Volume 3, Recent Advances in Engineering Science. Boston: Scientific Publishers; 1977. p 25–32, Proceedings, 10th meeting, Society of Engineering Science, NC: Raleigh; 1973.
105. Carter DR, Hayes WC. Bone compressive strength: the influence of density and strain rate. Science 1976;194: 1174–1176.
106. Dyson ED, Jackson CK, Whitehouse WJ. Scanning electron microscope studies of human trabecular bone. Nature (London) 1970;225:957–959.
107. Williams JL, Lewis JL. Properties and an anisotropic model of cancellous bone from the proximal tibial epiphysis. J Biomech Eng 1982;104:50–56.
108. Kaplan S, Hayes WC, Stone JL, Beaupre GS. Tensile strength of bovine trabecular bone. J Biomech 1985;18: 723–727.
109. Carter DR, Schwab GH, Spengler DM. Tensile fracture of cancellous bone. Acta Orthop, Scand 1980;51:733–741.
110. Gibson LJ. The mechanical behaviour of cancellous bone. J Biomech 1985;18:317–328.
111. Carter DR, Hayes WC. The compressive behaviour of bone as a two-phase porous structure. J Bone Jnt Surg 1977;59A:954–962.
112. Gibson LJ, Ashby MF. The mechanics of three-dimensional cellular materials. Proc R Soc London 1982;A382:25–42.
113. Galilei G, Discorsi E. Dimostrazioni Matematiche intorna a due nuove Scienze. 1638, Translated by H Crew, A deSalvio, editors. New York: Macmillan; pp 158–172. 1914. p 118–134.
114. Wolff J. Das Gesetz der Transformation der Knochen. Berlin: Hirschwald; 1892.
115. Frost HM. Bone remodelling and its relation to metabolic bone diseases. Springfield, IL: C Thomas; 1973.
116. Frost HM. Bone modelling and skeletal modelling errors. Springfield, IL: C Thomas; 1973.
117. Nathan PA, Fowler A. Remodeling of a metacarpal bone graft in a child. J Bone Jnt Surg 1976;58A:719–722.
118. Muller ME, Allgouer M, Willeneger H. Manual of Internal Fixation—Technique Recommended by the AO Group. Springer Verlag; 1970.
119. Pauwels F. Die Bedeutung des Bauprinzipien des Stütz, und Bewegensapparatens für, die Beanspruchung der Rohrenknochen. Z Anat Entwicklungs 1948;114:129–166.

120. Amtmann E. The distribution of the breaking strength in the femur. J Biomech 1968;1:271–277.
121. Seirig A, Kempko W. Behavior of *in vivo* bone under cyclic loading. J Biomech 1969;2:455–461.
122. Dietrick JE, Whedon G, Shorr E. Effects of immobilization upon various metabolic and physiological functions of bone. Am Jnl Med 1948;4:3–36.
123. Mack PB, La Chance PL. Effects of recumbency and space flight on bone density. Am J Clin Nutrition 1967;20:194–205.
124. Morey ER, Baylink DK. Inhibition of bone formation during space flight. Science 1978;201:1138–1141.
125. Wunder CC, Briney SR, Skangstad CA. Growth of mouse femurs during chronic centrifugation. Nature (London) 1960;188:151–152.
126. Wunder CC, Cook RM, Welch RC, Glade R, Fleming BP. Femur bending properties as influenced by gravity: I. ultimate load and moment for 3-G rats. Aviat Space Environ Med 1977;48:339–346.
127. Glucksmann A. Studies of bone mechanics *in vitro* I-Influence of pressure on orientation of structure. Anat Rec 1938;72:97–115.
128. Goodship AE, Lanyon LE, McFie M. Functional adaptation of bone to increased stress. J Bone Jnt Surg 1979;61A:539–546.
129. Hall BK. Developmental and cellular skeletal biology. New York: Academic Press, 1978.
130. Sumner-Smith G. Bone in clinical orthopaedics. New York: W. B. Saunders; 1982.
131. Lanyon LE, Magee PT, Bagott DG. The relationship of the functional stress and strain to the process of bone remodelling: an experimental study on the sheep radius. J Biomech 1979;12:593–600.
132. Woo SLY, Kuei SC, Amiel D, Gomez MA, Hayes WC, White FC, Akeson WH. The effect of prolonged physical training on the properties of long bone: a study of Wolff's law. J Bone Jnt Surg 1981;63A:780–787.
133. Liskova M, Hert J. Reaction of bone to mechanical stimuli, part 2, periosteal and endosteal reaction of the tibial diaphysis in rabbit to intermittent loading. Folia Morpholog 1971;19:310–317.
134. Wright KWJ, Yettram AL. An analytical investigation into possible mechanical causes of bone remodelling. J Biomed Eng (England) 1979;1:41–49.
135. Frost HM. Osteocyte death in vivo. J Bone Jnt Surg 1960;42A:138–143.
136. Cowin SC, Hegedus DH. Bone Remodeling I: Theory of adaptive elasticity. J Elasticity 1976;6:313–326.
137. Bassett CAL, Becker RO. Generation of electric potentials in bone in response to mechanical stress. Science 1962;137: 1063–1064.
138. Spadaro JA. Electrically stimulated bone growth in animals and man. Clin Orthopaed 1977;122:325–332.
139. Tischendorf F. Das Verhalten der Haversschen Systeme bei Belastung. Arch Entwicklungsmech Org 1951;145:318–332.
140. Piekarski K, Munro M. Transport mechanism operating between blood supply and osteocytes in long bones. Nature (London) 1977;269:80–82.
141. Jendrucko RJ, Hyman WA, Newell PH, Chakraborty BK. Theoretical evidence for the generation of high pressure in bone cells. J Biomech 1976;9:87–91.
142. Otter T, Salman ED. Hydrostatic pressure reversibly blocks membrane control of ciliary motion in paramecium. Science 1979;206:358–361.
143. Lakes RS, Katz JL. Viscoelastic properties and behavior of cortical bone, Part II, relaxation mechanisms. J Biomech 1979;12:689–698.
144. Justus R, Luft JH. A mechanicochemical hypothesis for bone remodelling induced by mechanical stress. Calc Tiss Res 1970;5:222–235.
145. Jendrucko RJ. Energetics of hydroxyapatite nucleation in bone, Proc 30th ACEMB, Los Angeles: 1977.
146. Martin RB, Burr DB. A hypothetical mechanism for the stimulation of osteonal remodelling by fatigue damage. J Biomech 1982;15:137–139.
147. Rasmussen H, Bordier P. The Physiological and Cellular Basis of Metabolic Bone Disease. Williams & Wilkins; 1974.
148. Davidovitch Z, Furst L, Shanfield JL, Montgomery PC, Kelischeck S, Laster L, Korostoff E. Biochemical mediators of electrical stimulation of bone cells. 35th ACEMB. Philadelphia: Sept. 1982. p 217.
149. Rodan GA, Bourret LA, Norton LA. DNA synthesis in cartilage cells is stimulated by oscillating electric fields. Science 1978;199:690–692.
150. Sutherland J, Rall M. The relation of adenosine 3′:5′-phosphate and phosphorylase to the actions of catecholamines and other hormones. Pharmacol Rev 1977;12:265–299.
151. Fukada E, Yasuda I. On the piezoelectric effect of bone. J Phys Soc J 1957;12:1158–1162.
152. Reinish G. Piezoelectric properties of bone as functions of moisture content. Nature (London) 1975;253:626–627.
153. Lakes RS, Katz JL. Dielectric relaxation in cortical bone. J Appl Phys 1977;48:808–811.
154. Burr AJ. Measurements of the dynamic piezoelectric properties of bone as a function of temperature and humidity. J Biomech 1976;1:495–507.
155. Anderson JC, Eriksson C. Electrical properties of wet collagen. Nature (London) 1968;218:167–169.
156. Anderson JC, Eriksson C. Piezoelectric properties of dry and wet bone. Nature (London) 1970;227:491–492.
157. Williams WS. Sources of piezoelectricity in tendon and bone. CRC Crit Rev Bioeng 1974;2:95–117.
158. Lakes RS. The role of gradient effects in the piezoelectricity of bone. IEEE Trans Biomed Eng 1980;BME27:282–283.
159. Korostoff E. Stress generated potentials in bone: relationship to piezoelectricity of collagen. J Biomech 1979;10:41–44.
160. Korostoff E. A Linear piezoelectric Model for characterizing stress generated potentials in bone. J Biomech 1979;12:335–347.
161. Gjelsvik A. Bone remodeling and piezoelectricity II. J Biomech 1973;6:187–193.
162. Guzelsu N. A piezoelectric model for dry bone and tissue. J Biomech 1978;11:257–267.
163. Johnson M, Chakkalakal D, Harper RA, Katz JL. Comparison of the electromechanical effects in wet and dry bone. J Biomech 1980;13:437–442.
164. Chakkalakal DA, Johnson MW, Harper RA, Katz JL. Dielectric properties of fluid saturated bone. IEEE Trans Biomed Eng 1980;BME-27:95–100.
165. Athenstaedt H. Permanent electric polarization and pyroelectric behavior of the vertebrate skeleton. VI, the appendicular skeleton of man. Z Anat Entwickl Gesch 1970;131:21–30.
166. Lang SB. Pyroelectric effect in bone and tendon. Nature (London) 1966;212:704–705.
167. Bassett CAL, Pilla AA, Pawluk RJ. A non-operative salvage of surgically resistant pseudarthrosis and non-unions by pulsing electromagnetic fields: a preliminary report. Clinical Orthop 1977;124:128–143.
168. Brighton CT, Friedenberg ZB, Mitchell EI, Booth RE. Treatment of non-union with constant direct current. Clinical Orthop 1977;124:106–123.

169. McLeod KJ, Rubin CT. The effect of low frequency electrical fields on osteoporosis. J Bone Joint Surg 1992;74-A:920–929.
170. Skerry TM, Pead MJ, Lanyon LE. Modulation of bone loss during disuse by pulsed electromagnetic fields. J Orthop Res 1991;9:600–608.
171. Granhed H, Jonson R, Hansson T. The loads on the lumbar spine during extreme weight lifting., Spine 1987;12(2): 146–149.

Reference List

Gilmore RS, Katz JL. Elastic properties of apatites. J Mat Sci 1982;17:1131–1141.
Katz JL, Ukraincik K. On the anisotropic elastic properties of hydroxyapatite. J Biomech 1971;4:221–227.
Lakes RS, Katz JL. Viscoelastic properties of bone. In Hastings G, Ducheyne P, editors. Natural and Living Biomaterials. Washington, (DC): CRC Press; 1984.
Ortmann R, Perkins JP. Stimulation of adenosine 3′:5′ monophospate formation by prostaglandins in human astrocytoma cells. J Biol Chem 1977;252:6019–6025.

See also BIOMATERIALS FOR DENTISTRY; BONE CEMENT, ACRYLIC; TOOTH AND JAW, BIOMECHANICS OF.

BONE CEMENT, ACRYLIC

CHAODI LI
YAN ZHOU
University of Notre Dame
Notre Dame, Indiana

INTRODUCTION

Many years of intensive research by Rohm led to the development of poly(methyl methacrylate) (PMMA), the basis of bone cement, in 1934 (1,2). This polymer was reportedly first used to close cranial defects in monkeys in a medical application in the late 1930s (2). The use of acrylic bone cement in orthopedic surgery was first advocated in 1951 by Kiaer and Jansen. It was applied as pure anchoring material by fixing acrylic glass caps on the femoral head after removing the cartilage (2–4). At that time, the femoral components in total hip replacements were still implanted by simply press-fitting the prosthesis tightly into the prepared intramedullary (marrow) canal of the femur. Patients were confined to bed for a relatively long period of time after surgery and often experienced pain later from loosening of the implant. In 1958, Sir John Charnley first applied the self-curing bone cement for the fixation of artificial joints (4). In this surgery, the cement filled the free space between the prosthesis and the bone (2). Bone cement served as a mechanical interlock between the metallic prosthesis and the bone and it has been found to be an appropriate material to transfer the load consistently (5). In 1969, bone cement was approved for general use within the United States and since then the number of total hip and knee replacements has increased dramatically (3). Currently, it is the only material used for anchoring cemented arthroplasties to the contiguous bones (6). More than one-half million hip replacement surgeries are performed every year worldwide and 70% of the surgeries are performed using bone cements (7–10).

Bone cement is also extensively used in the fixation of pathological fractures, spinal surgery, and neurosurgery (11). Every year, osteoporosis results in $>310{,}000$ fractures in the United Kingdom alone (12). Vertebral compression fractures occur in 20% of people over the age of 70 years and in 16% of postmenopausal women (12). Although traditional conservative techniques, such as bed rest and the prescription of analgesics may be successful in a proportion of cases, a significant number of sufferers remain in long-term pain. Vertebroplasty is now being used extensively for vertebral compression fracture treatments (12,13). This technique entails the percutaneous injection of bone cement into the fractured vertebra in attempts to stabilize the fracture and reduce pain. Studies have reported excellent pain relief and improved function in most patients (12,13). Bone cements have also been applied as an adjunct to internal fixation for treating fractures. Bone cement fills voids in bone, thereby reducing the need for bone grafts, and may improve the holding strength around the devices in osteoporotic bone. Recently, the technique of PMMA bone cement injection into the osteoporotic proximal femur was proposed (14). Results indicated that cement injection increased the peak fracture load >80 and 20% in the simulated fall and one single limb stance configurations, respectively (14). This technique could become a treatment option to solve the problems with osteoporotic hip fractures in patients at risk. Such treatment may hold a significant impact on the economy and human health as hip fractures result in $\sim 300{,}000$ hospital admissions annually in the United States with an estimated $9 billion in direct medical costs (14). Therefore, bone cements are significantly valuable for biomedical applications.

In many cases, the bone cemented implants perform satisfactorily for years. However, failure of the implants may be linked to bone cement properties. It is well recognized that there are a number of drawbacks that exist with the usage of bone cement, significant ones including its poor mechanical properties and the potential necrosis of bone tissues (6,15,16). For example, bone cement has been called the "weak link" in the prosthesis system and long-term loosening of the prosthesis has been attributed to its mechanical disintegration (10). It is worth pointing out that aseptic loosening of a cemented arthroplasty is a multifactorial phenomenon involving interfacial failure, bone failure, bone remodeling, and cement failure (6). It is not firmly established whether the drawbacks of bone cements contribute to the initiation or are the consequence of aseptic loosening of the implant (6). In spite of the many drawbacks of bone cement, the survival probabilities of recently cemented joint replacements are still high. Currently, cemented total hip arthroplasty shows survival rates of 90% at 15 years and 80–85% at 20 years (10,15,17–21). Increasing bone cement properties along with improving cementing techniques and implantation methods may further extend cemented implant longevity.

In recognizing the drawbacks of bone cement usage, there have been considerable efforts to secure the implants without using cement. Early noncemented prostheses were the press-fitted or screwed-in types (10). The press-fit

techniques must lead to good initial fixation; otherwise, mobility will occur and this will facilitate wear and/or formation of fibrous tissue. Such techniques have limited successes due to a number of reasons, including bone resorption, geometrical constraints, and increased operational fractures. Noncemented porous coated prostheses were also introduced to provoke bony ingrowth for improved fixation. The noncemented prostheses are used predominantly in younger patients (< 60 years), where it is assumed that these prostheses eventually simplify a revision operation (10). Bone ingrowth strategies are not appropriate for older patients. Research indicates that some of the noncemented devices have failed, but others are doing well in the mid- to long- term (10). The clinical applications of cementless total hip and knee arthroplasties induce a new set of problems, including perioprosthetic osteolysis, high thigh pain, and failure of the bone–implant interface (6). Thus, there is currently a resurgence of interest in bone cement (6). Definite conclusions about the ultimate clinical performances of cement or cementless fixation devices require longer term studies (10). The debate is still open as to which fixation method is best. The development of cementless modes of fixation is an area of active research and clinical practice; however, acrylic bone cement continues to be the most commonly used nonmetallic implant material by orthopedic surgeons (6,11).

The literature on bone cement is voluminous with respect to its material development, manufacturing, thermal response, chemical and biological effects to bone, and its short- and long-term physical and mechanical properties (static, fatigue, etc.). This article discusses some of these aspects, especially on cement's material characteristics. For more ample assessment, the readers should refer to other excellent comprehensive reviews [e.g., Krause (3,11), Lewis (6,22), Kuhn (2), Saha (23), Kenny (5), Deb (24), Hasenwinkel (15), Serbetci and Hasirci (16)]. The article is arranged into several sections. It begins with the description of bone cement compositions, followed by a section on cement setting procedure and cementing technique. Then, the thermal and volumetric change effects are discussed. The final section focuses on the physical and mechanical properties of cement. Finally, the article ends with a brief summary.

CEMENT COMPOSITIONS

There are a number of bone cements [> 60 types (2)] available to the orthopedic community. Some popular cements currently available in the United States are given in Table 1 (2,6,11). Most of the present commercial bone cements have similar compositions (Table 2) (2,11,25). The presently used form of bone cement is predominantly the same as the one Sir John Charnley introduced. All acrylic bone cements on the market are chemically based on the identical basic substance: methyl methacrylate (MMA) (2). Pure MMA exhibits a shrinkage of ~21% during polymerization (2), and the polymerization temperature can increase to 100–120°C. Such a high shrinkage is intolerable for use in bone cement (2). For this reason, bone cements are offered as two-component systems in the marketplace: a prepolymerized powder and a liquid mm monomer (2). The MMA in aqueous suspension is prepolymerized in easily cooled reaction boilers. The polymer, obtained in the form of tiny balls (< 150 μm), is easily dissolved in the MMA. By using the prepolymerized polymer powder, both the shrinkage of the sample and the temperature of the reaction can be considerably decreased. In most bone cements on the market, the mixing ratio is two to three parts powder to one part monomer. This reduces the shrinkage and the generation of heat by at least two-thirds, as only the monomer is responsible for these reaction symptoms (2).

Usually, the solid part of bone cement consists of prepolymerized PMMA beads ranging in size from 1 to 150 μm. Other kinds of polymers sometimes are added, including poly(ethyl acrylate), poly(methyl acrylate), poly (styrene), and poly(butyl methacrylate). Free radicals of benzoyl peroxide (BPO) are present within the beads as remnants from the emulsion polymerization process (the process by which most of the beads are manufactured). An additional amount of BPO is mixed with the solid to obtain 1–2.5% by weight benzoyl peroxide. In addition, bone cements generally contain ~10–15% by weight barium sulfate, zirconia, or other additive. The presence of barium sulfate or zirconium dioxide in the powder is necessary for

Table 1. List of Popular Commercial Bone Cements Currently Available in the United States[a]

Bone Cements	Manufacturer or Distributor
CMW1	Depuy, Warsaw, IN
CMW3	Depuy, Warsaw, IN
Palacos R	Smith and Nephew, Memphis, TN
Simplex P	Stryker Howmedica Osteonics, Rutherford, NJ
Osteobond	Zimmer, Warsaw, IN
Zimmer dough-type	Zimmer, Warsaw, IN

[a] See Refs. 2,6, and 11.

Table 2. Compositions of Commercial Bone Cements Currently Available in the United States[a,b]

Liquid component	
Methylmethcrylate (monomer)/ Butylmethacrylate (binding agent)	~98
Activator/Co-initiator: Dimethyl-paratoluidine	0.4–2.75
Stabilizer/inhibitor/radical catcher: Hydroquinone, Ascorbic acid	15–75 ppm
Coloring: Chlorophyll	267 ppm
Powder	
Poly(methyl methacrylate)/copolymer	~90
Initiator: Benzoyl peroxide	0.5–3
Opacifier: $BaSO_4$ or ZrO_2	10–15%
Coloring: Chlorophyllin	200 ppm
Antibiotics: Gentamicin, erythromycin, and colistin	

[a] Compositions are in percent (w/w) except where stated otherwise.
[b] See Refs. 2,6,11 and 25.

a clinical reason. The original bone cements, which did not contain radiopacifiers could not be visualized on radiographs. The main components in the liquid phase are MMA, and, in some bone cements, other esters of acrylic acid or methacrylic acid, one or more amines (as activators for the formation of radicals), a stabilizer and, possibly, a colorant (2). The amine in the bone cement, *N*,*N*-dimethyl-*p*-touluidine (DMPT), acts as an accelerator. The liquid component also consists of 50–100 ppm hydroquinone, which inhibits the polymerization reaction within the monomer and allows for storage of the liquid component. Some cement also contains chlorophyll that gives it a green color. This allows better distinction from body tissues during surgery.

Prosthesis-related infection is described as a devastating failure scenario of a cemented orthopedic implant. Infectious complications of a cemented prosthesis lead to a deterioration of function and increase pain. Buchholz and Engelbrecht first reported on the possibilities of mixing antibiotics in bone cement in 1970 (17). They considered gentamicin sulfate to be the antibiotic of choice because of its wide-spectrum antimicrobial activity, its excellent water solubility, its thermal stability and its low allergenicity. Apart from gentamicin, other antibiotics have also been used as an additive to bone cement. The combination of erythromycin and colistin is an example that made it to a commercial product (17). In most cases, the manufacturers make their antibiotic cements by simply mixing antibiotic to a plain cement version they have (2,15,16).

Currently, bone cement fracture is regarded as a major factor in the mechanical failure of implant fixation (26–28). It is directly related to the mechanical properties of the cement, especially the resistance to fracture of the cement in the mantle at the cement–prosthesis interface or the cement–bone interface. Thus, many investigators have attempted to incorporate second phase materials including polyethylene (29), hydroxyapatite (30), PMMA (31), Kevlar (32,33), carbon (34–36), titanium (37), and steel (38–40) to improve the fatigue properties and fracture toughness of the PMMA. Most of the results regarding the properties of these composite materials have been encouraging; however, the biocompatibility issues regarding some of these fibers are as yet unresolved (6). Consequently, none of the commercial bone cements have incorporated these fibers in cements on the market.

CEMENT SETTING AND CEMENTING TECHNIQUE

Most structural materials are fabricated under controlled conditions at a factory, then transported and assembled on site. Bone cement is one of the few structural materials that are created *in situ*. The surgeon prepares the bone cement directly at the operation table according to the manufacturer's instructions. All of the cements are supplied in sets of the polymer powder and the monomer liquid components packed in two separate containers within a package. At the time of surgery, the liquid monomer and powder are mixed, the DMPT reacts with the BPO to generate free radicals, which in turn are used in the additional polymerization of the MMA monomers to form PMMA. Polymer chains from the PMMA become available for free radical polymerization and entanglements of these chains with newly formed chains lead to an intimate connection between the newly formed PMMA with what was already present. The resulting product is a doughy mixture that later polymerizes to a hard and brittle substance.

Table 3. Four Phases of Bone Cement Polymerization Process[a]

Phases	Time Duration, min	Characteristics
I. Mixing	1–2	Wetting Cement relatively liquid (low viscous)
II. Waiting	2–3	Swelling + polymerization Increase of viscosity Polymer chains, less movable Sticky dough
III. Working	5–8	Chain propagation Reduced movability Increase of viscosity Heat generation
IV. Setting	2–6	Chain growth finished No movability Cement hardened High temperature

[a]See Ref. 2.

The curing process of acrylic PMMA bone cement can be divided into four basis steps (2): the mixing, waiting, working, and hardening phase. The characteristics of these phases, well described by Kuhn (2), are shown in Table 3. The time at which the cement does not stick to the surgeon's glove is referred to as dough time (Fig. 1). The waiting phase ends at this time point. This occurs $\sim$2–3 min after the beginning of mixing for most PMMA cements in an ambient temperature of 23 °C (2,11). The working phase is the time during which the surgeon can easily apply cement to the femur. For manual application, the cement must no longer be sticky during this phase, and the viscosity must not to be too high. With the use of mixing systems, the user needs not to wait until the cement is no long sticky. The working time from the end of dough time to the cement is too stiff to manipulate is usually 5–8 min. The cement will fully polymerize to a hardened mass within 8–12 min after initial mixing (2,11).

The quality of the cement dough produced in the operation room will have considerable influence on the clinical long-term result of a cemented prosthesis. Mixing of cements is an important step, as it has a noticeable influence on the mechanical properties of acrylic bone cements (2,6,24). In the 1970s, loosening of the femoral stem was the most common reason for total hip arthroplasy revision, often occurring 5–10 years postoperatively with early component design and cement technique (21). Early methods of cement preparation involved hand mixing in open air of the MMA monomer with the prepolymerized powder mixture. A slurry was formed that could be hand patted or injected into the femoral canal. Many air bubble voids were

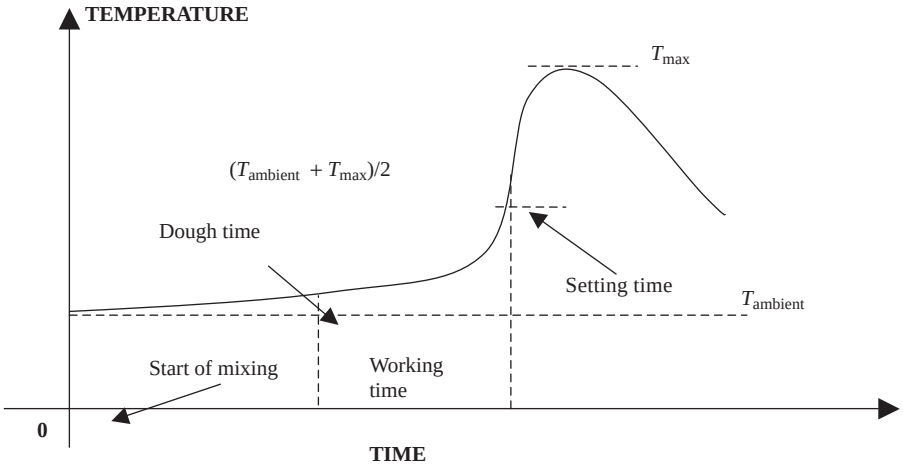


Figure 1. A typical temperature changes with respect to time during cement polymerization. Time zero is designated when the monomer liquid is added to the polymer powder.

created in the mixture during the hand-mixing process. To address this problem in cemented arthroplasty, intensive studies have attempted to improve the technique of cemented stem insertion (21). This results in advancing the cementing techniques from first generation cementing to second and third generation cementing (Table 4). Traditionally, bone cement was mixed using a spatula–bowl arrangement (first generation), which can have the consequence of introducing a high degree of porosity into the cement structure. In addition, the person mixing the cement was exposed to a high level of methyl methacrylate vapors. Second-generation cementing mainly consisted of the use of a canal plug, a cement gun and a high strength cement (21,41,42). Injection guns have been advocated for application of bone cement because long slender injection tips are useful for inserting the cement deep into the cavity, instead of trying to apply it by hand. Gun application also reduces the tendency to form laminations and voids in the cement and also reduces the inclusion of blood into the cement. Substantial improvements in stem survival were reported using such techniques, resulting in stem loosening rates of $< 5\%$ at 5–10 years (21,42). The introduction of porosity reducing measures marked the start of third generation cementing techniques. Cracks may initially occur at the voids in bone cements, which act as stress concentration points (24). Vacuum mixing is a typical ways to achieving pore reduction. Other third-generation developments in cementing techniques are the use of prosthesis positioning devices that ensure correct placement of the prosthesis, pressurization, and enhanced surface finishes (21,43). Attempts to improve cement-bone interlock using techniques such as endosteal preparation, retrograde cement insertion and cement pressurization improve implant survival. Improving cementing techniques have assisted to dramatically decrease the number of failure in the last three decades. Currently, cemented total hip arthroplasty shows survival rates of 90% at 15 years and 80–85% at 20 years (10,17,18,21).

Table 4. Developments in Cementing Techniques

Generations	Techniques	Time
1st Generation	Finger packing No cement restrictor	1960s
2nd Generation	Intramedullary femoral plug Cement gun High-strength cement	1970s
3rd Generation	Pressurization of cement after insertion Porosity reduction (vacuum mixing) Surface roughening or texturing Precoating Avoiding trochanteric osteotomy	mid-1980s–

POLYMERIZATION HEAT

Bone cement generates significant thermal energy during cement setting (an energy of 52 $kJ \cdot mol^{-1}$ of MMA (2)) and this may result in a temperature increase in the cemented system. The exothermic response of the bone cement can be characterized according to American Society for Testing and material (ASTM) standard F451 (44) or international standard ISO 5833. In the ASTM test, a package of cement is mixed at an air conditioned room ($23 \pm 1\,^{\circ}C$, $50 \pm 10\%$ relative humidity) as directed by the manufacturer's instructions. Within 1 min after doughing time, ~ 25 g of the dough is then gently packed into an ASTM specified test mold to achieve a 60 mm diameter with 6 mm thick disk cement mantle (44). The setting curve of temperature change with respect to time (Fig. 1) is recorded with a thermocouple (usually No. 24 gage wire k-type thermocouple) placed at the cement mantle center from the onset of mixing until cooling is observed. The setting time is defined as the time from initial mixing to the time at which the temperature of the polymerising mass has reached half of the maximum temperature (peak temperature) (44). The setting times range from 6 to 14 min and peak temperatures range from 54 to $83\,^{\circ}C$ for the popular commercial bone cements (2,11). A number of variables will affect the measured setting time and peak temperature, including the powder/liquid ratio of the cement, cement preparing method, cement mantle thickness, the initial temperature of the cement, and the ambient conditions (11). By mixing bone cement with Howmedica Mix-Kit I system and the Zimmer Osteobond vacuum system, Dunne and Orr (45)

found peak temperatures increased from 36 to 46 °C and 41 to 59 °C for Palacos R and CMW3 bone cement, specifically. Cold cement in a cold room takes longer to set. Precooling cement prior to use extends the cure time and decreases the peak temperature (46). Using a test mold and cementing techniques that simulated a clinical situation, Iesaka et al. (47) showed that the peak temperatures at the bone–cement interface were 53.1, 50.2, and 48.8 °C, while the polymerization time was 4.1, 8.3 and 11.2 min, when the monomer component of bone cement was initially at 37, 23, and 4 °C, respectively. Research also found that bone cement thickness has significant effects on the thermal responses. Meyer et al. (48) found a setting temperature of 60 °C with 3 mm thick specimens of Simplex R, and 107 °C with 10 mm thick specimens. Sih and Connelly (49) showed that the temperatures were 41, 56, and 60 °C for cement thicknesses of 1, 5, and 6-7 mm, respectively. Vallo (50) and Li et al. (51) demonstrated similar results with finite element modeling. Test mold materials also affect the measured setting time and peak temperature (50,51). According to the ASTM methodology, there is no limit to the setting time. However, the maximum allowable temperature for bone cements is 90 °C (44). There are limits stipulated for setting time and doughing time with ISO 5833 Standard.

The importance of temperature rise is that it may result in thermal necrosis of the bone tissue surrounding the implant (2,11,52–56). *In vitro* studies have shown that maximum temperature of bone cement can be reach higher than 100 °C, varied between 37 and 122 °C in different reports (57). Wang et al. (58) measured four different brands of bone cement (Palaces R, Simplex P, Sulfix, and CMW 1) during polymerization and the peak temperature in the cement was 46–124 °C. Clinical tests show a considerable lower temperature in the body (2,55). The bone-cement interface temperatures have ranged from 35 to 70 °C (2,11,52,3,54,55,57,59). Reasons for this observation are the thin layer of cement (~3–5 mm) and the blood circulation and heat dissipation in the vital tissue connected with it (2). Moreover, further heat dissipation of the system is attained via the prosthesis (2). Belkoff and Molloy (59) found that peak temperatures at the anterior cortex ranged from 44 to 113 °C, in the center ranged from 49 to 112 °C, and 39 to 57 °C at the spinal canal in the *ex vivo* vertebroplasaty tests. Toksvig-Larsen (57) performed tests with 31 total hip replacements and showed that the maximum temperature in the acetabulum ranged from 38 to 52 °C and in the femur from 29 to 56 °C. It has been found that not only the temperature, but also the exposure time plays a significant role in direct thermal cell necrosis (52,60–63). The findings of Moritz and Henriques (60) suggest that epithelial cell necrosis occurs 30 s after exposure to a temperature of 55 °C and 5 h after exposure to 45 °C. Lundskog (61) roughly confirmed these trends for bone cells, although he found a slightly lower threshold level. For example, he observed bone cell necrosis after 30 s at 50 °C. He also established that the regenerative capacity of the bone tissue is only damaged after exposure to temperatures of 70 °C and above. Eriksson and Alberksson (64) found the temperature threshold for impaired bone regeneration to be in the range of 44–47 °C for 1 min exposure. Thermal damage to bone tissue caused by cement polymerization cannot be ruled out and attempts to lower the potential degree of thermal necrosis have been investigated. To reduce the risk of thermal injury it is necessary to avoid exposing bone to temperatures above a certain threshold (62). The thermal responses rely on the quantity of heat produced by the bone cement (cement formulations and cement volume, etc.), rate of heat produced and how the heat conducts (55). The temperature peak can only by influenced to a small extent by adding heat-conducing radiopaque media or by slightly changing the chemical composition of the liquid (2). This, will, however, result in quite different dissolution properties with the polymer, which means it will result in different working properties and, usually, a significant reduction in mechanical stability (2). The thicker the cement mantle, the greater the volume of material and hence the more heat generated. However, simply reducing the thickness of the mantle is not a favorable option as the mechanics of the joint is affected by this. A slightly reduced level of heat generation has been shown by vacuum mixing the bone cement (45,57). Precooling the cement constituents prior to mixing increases the setting time, but has only minor effects on the maximum temperature of the cement (56). The use of a precooled femoral prosthesis did not affect the peak temperature as well (56,57). To avoid local tissue damage, surgeons may irrigate the implant with ice-cold saline during the polymerization process to decrease both the duration and the level of temperature elevation. Without cooling, the temperature was 49 °C (41–67) at the bone–cement interface, while it decreased to 41 °C (37–47) with water cooling in 19 cases of arthroplasties (65). Investigations on the potential tissue thermal necrosis have obtained varied results: while thermal bone necrosis due to cement curing heat has not been widely reported, some studies showed that thermal necrosis of bone did occur (52,54,59). Nevertheless, few would disagree that a clear understanding of the potential thermal necrosis problem requires further investigations.

VISCOSITY

During the working stage of the cement, its viscosity must be low enough to make it easy to force the cement dough through the delivery system and cause it to flow and penetrate into the interstices of the bone surface in a very short time (6,11,22). *In vitro* determination of viscosity is usually accomplished with the use of a capillary extrusion or rotational rheometer (6,11,22). The viscosity of the material can be determined by causing the material to flow at a specific shear rate and measuring the shear stress of the fluid against the stationary instrument (11). Viscosity varies across cements. The dynamic viscosity of its dough during the mixing period has been used to categorize bone cement materials into low viscosity brands (e.g., Osteopal), medium-viscosity brands (e.g., Simplex P), and high viscosity brands (e.g., Palacos R) (22). High viscosity bone cements typically have a doughy consistency; low viscosity cements are similar to viscous oil in consistency. Low viscosity cements have a long liquid

phase, or low viscosity wetting phase. The cement remains sticky for quite some time. Viscosity increases rapidly during the working phase, and the doughy cement becomes warm and sets quickly. High and low viscosity bone cements have different handling characteristics and require different cementing techniques. For optimal results, it is necessary to observe the specific mixing instructions for a given bone cement in combination with a given mixing system. Although some researchers suggest that low viscosity brands may have longer fatigue lives compared to high viscosity ones, some report no significant difference (22).

POROSITY, VOLUMETRIC CHANGES, AND RESIDUAL STRESS

Polymerized bone cement is a porous material, containing macropores (pore diameter > 1 mm) and micropores (pore diameter ≈ 0.1–1 mm) (6). The degree of porosity varies with cement brands and mixing methods. Jasty et al. (66) found that the porosities were 11.99% (CMW), 9.70% (Palacos R), 9.39% (Simplex P), 12.38% (Zimmer Regular), and 5.00% (Zimmer LVC) using manual mixing, while they were 6.00% (CMW), 4.25% (Simplex P), 6.00% (Zimmer Regular) and 4.15% (Zimmer LVC) with centrifugation. For CMW 1 cement, Muller et al. (67) found that the porosity decreased from 6.67 to 1.28% when using vacuum-mixing instead of hand-mixed methods. The pores generated in polymerized bone cement have been attributed to several sources (6,68–72). It may result from the air that is initially present in the powder interstices; entrapped during blending, mixing, transfer, or delivery; or entrained during insertion of the metal stem. Evaporation of the liquid monomer at the high temperatures of polymerization may contribute. Another aspect of porosity development is polymerization shrinkage.

The presence of pores in the polymerized cement may affect its mechanical properties. Pores may act as stress risers and initiation sites for cracks, rendering the cement susceptible to early fatigue failure (6). However, pores also may play a role in blunting crack propagation, thereby prolonging the life of the implant (73). *In vitro* experiments, the static and fatigue strength of bone cement both usually decrease with increased porosity (6,22,23,69,70). Reducing the porosity of both bulk cement and its interfaces should be of clinical benefit. Several methods of reducing the porosity of the bone cement have been developed (66). One commonly used method of void reduction is the centrifugation of a chilled, premixed bone cement prior to insertion into the bone. The other one is the mixing of the bone cement in a vacuum environment. Both techniques have been shown to substantially reduce the porosity of bone cement. It has been observed that mixing procedures plays a significant role in determining the quality of bone cement produced (2,74,75). The influence of vacuum mixing on the pores results in a 15–30% improvement of the bending strength of Palacos R while centrifuging Simplex P cement reduced its porosity from 9.4 to 2.9%. Experiments have shown that relative to hand mixing, centrifugation or vacuum mixing leads to a substantial reduction in porosity (2,6). The extent of such a reduction depends on many mixing variables, including mixing system, monomer storage temperature, vacuum pressure and centrifugation speed, and durations. In the hand-mixing process, air bubbles are mixed into the dough by thorough mixing. The porosity of the material is high and mechanical stability is endangered. Slower mixing of cement over a shorter time decreases air voids within cement and thus improves the strength characteristics. A high degree of porosity is found to exist in cement that is inadequately mixed (74). Monomer bubbles can easily appear, which may develop during the evaporation of the monomer while evacuating the system or later during polymerization under high pressure by the faulty use of vacuum-mixing systems.

Cement porosity distributions, especially at the bone–cement interface, may affect the cemented system. There is strong evidence that cracks in the cement are initiated at voids, particularly at the cement–prostheses interface (69). The preferential formation of voids at this site results from shrinkage during bone cement polymerization and the initiation of this process at the warmer bone–cement interface, which causes bone cement to shrink away from the prosthesis (2,68,69). It is expected that a reversal of polymerization direction would shrink the cement onto the prosthesis and reduce or eliminate the formation of voids at this interface (69). One innovative surgical approach to affect this behavior is to preheat the prosthesis prior to implantation. Results indicated that voids near the cement–prosthesis interface decreased significantly (69,76). The porosity at the cement–prosthesis decreased from 16.4 to 0.1% when the prosthesis was preheated to 37°C from room temperature, 23 °C. Additionally, the residual stress due to such polymerization curing was shown to decrease significantly at the cement–prosthesis interface (77). Studies also showed that preheating the prosthesis prior to implantation is unlikely to produce significant thermal damage to the bone when compared to implanting a prosthesis initially at room temperature (46,69). However, this procedure may induce more voids at the bone–cement interface, and its effects on the damage at this region should be studied (69,78).

The conversion of the monomer molecules into a polymer network is accompanied with a closer packing of the molecules, which leads to bulk contraction (68,79). A number of devices for determining the volumetric changes have been applied, including using a mercury dilatometer or a water displacement dilatometer (79,80). Theoretical calculations predict that bone cement polymerization will produce a volumetric shrinkage of 8% (81). Muller et al. (67) found volume shrinkages of CMW 1 bone cement were 3.43 and 5.99% with hand and vacuum mixing, respectively. Gilbert et al. (68) found shrinkages of Simplex P cement were 5.09 and 6.67%, respectively. The measured volume shrinkage is less than the theoretical prediction. This can be explained by void growth during polymerization (67).

In a situation where a curing material is bonded on all sides to rigid structures or constrained, bulk contraction cannot occur freely, and shrinkage must be compensated for by some kind of volume generation. This can come from a strain on the material and mainly for dislodgement of the bond, increase in porosity or internal loss of coherence (79).

Table 5. Comparison of the Values of Three Mechanical Properties of Six Different Bone Cements Under Same Test Regimes[a]

Cement Brands	ISO 5833 Bending Strength, MPa	Bending Modulus, MPa	Compressive Strength, MPa	DIN53435 Bending Strength, MPa	Impact Strength, $kJ \cdot m^{-2}$
CMW1	67.0	2634	94.4	86.2	3.7
CMW3	70.3	2764	96.3	72.4	2.9
Palacos	72.2	2628	79.6	87.4	7.5
Simplex P	67.1	2643	80.1	70.5	3.9
Osteobond	73.7	2828	104.6	80.1	3.5
Zimmer dough	62.5	2454	75.4	77.0	5.0

[a]See Ref. 2.

Shrinkage of the polymerizing cement *in vivo* (i. e., a constrained state) therefore might result in the development of porosity, both at the interfaces and inside the bulk cement. Thus, polymerization shrinkage may be a significant factor in porosity development (68). On the other hand, polymerization may induce high residual stresses. Shrinkage stress occurs when the material contraction is obstructed and the material is rigid enough to resist sufficient plastic flow to compensate for the original volume (79). The process of cement curing is a complex solidification phenomenon where transient stresses are generated and the residual stresses vary with different initial and boundary conditions during curing. A number of approaches have been used to estimate or measure the level of shrinkage stress in bone cement around femoral replacements, including theoretical model, finite element analysis, strain-gage methods and photoelastic methods (52,77,82–88). Currently, the subject of residual stress has often been neglected because it is assumed that residual stress will relax due to the viscoelastic properties of the cement. However, transient and residual stresses are believed to affect cement mechanical responses. Inclusion of the residual stress at the interface resulted in up to a four-fold increase in the von Mises cement stresses compared to the case without residual stresses (83,84). Recently, Orr et al. proposed that residual stresses are sufficient to initiate crack propagation in the cement before any load is applied (85). Lennon and Prendergarst have experimentally observed that residual stresses in the cement may induce cracking even before weight bearing of the cement (89). The initial residual stresses may have immediate effects influencing the possible initiation of cracks and debonding at the cement–prosthesis interface.

MECHANICAL PROPERTIES

Bone cement fills the space between the prosthesis and the bone; this connection is only a mechanical bond (1). Cement mechanical properties are therefore of particular significance for the performance of acrylic bone cement because cement must endure considerable stresses *in vivo*. There are many physical and mechanical properties of the cement that are considered germane to its clinical performance in the construct, including quasistatic tensile and compressive strength, modulus and ultimate strain, flexural strength and modulus, shear strength and modulus, fatigue properties (e.g., work of fracture, fracture toughness, fatigue resistance, fatigue crack propagation resistance), and creep (6). The mechanical properties of acrylic bone cement have been widely reported in the literature. Kuhn (2) investigated the static bending and compressive characteristics of a number of bone cements under the same test regimes (2). Test results of some of the most commonly used bone cement in United States are listed in Table 5. Osteobond cement was shown to have both the highest ISO 5833 bending strength and compressive strength among these six cement brands, while Zimmer dough has both the lowest ISO 5833 bending strength and compressive strength. All the measured bending strengths of the bone cement using the DIN53435 standard were larger than the ones using ISO 5833 standard, with Palacos R bone cement having the highest bending strength in this case. Harper and Bonfield (90) found a wide range of tensile strength (Table 6) and fatigue failure cycles (Table 7) results. They found that the Palacos R and Simplex P cements were significantly higher in tensile strength compared to the other cements tested with exception of CMW 3. There was no statistical difference between the values obtained for CMW 1 and Osteobond. The differences among the fatigue results for the different cements were much larger than those found with the static tensile results. The highest Weibull median fatigue cycles to failure obtained for Simplex P and Palacos R were considerably higher than found for Zimmer dough type. Harper and Bonfield (90) found that there was some correlation between the static and fatigue strengths, but the ranking of static strength does not exactly follow that of fatigue life. The fatigue results were found to correlated well to the clinical data (91): the order of success of implants with the cement brand was the same as that obtained from the fatigue test.

It has long been recognized that PMMA surgical bone cement undergoes viscoelastic (creep) deformation under physiological loads (92,93). Many studies have been performed to assess the viscoelastic properties of bone cement (6,23). Radiological observations of hip stems have shown subsidence of the stem within the cement mantle without visible cement fractures (94). Creep has been implicated in prosthesis subsidence, in particular subsidence of the femoral stem in hip replacements (6,95). Excessive subsidence can lead to prostheses loosening. Lu and McKellop (92) studied the effects of cement creep on the subsidence of

Table 6. Comparison of Tensile of Six Different Bone Cements[a]

Cement Brands	Ultimate Strength, MPa	Modulus of Elasticity, GPa	Strain at Fracture, %
CMW1	39.1	2.96	1.60
CMW3	44.7	3.53	1.36
Palacos	51.4	3.21	2.25
Simplex P	50.1	3.43	1.87
Osteobond	38.2	3.38	1.41
Zimmer dough	31.7	2.79	1.43

[a]See Ref. 90.

the stem and on the stress within the cement using a cyclic load and three interface bonding conditions (bonded, frictional, and debonded). Results showed that the creep deformation of the cement was accompanied by additional subsidence of the stem and a decrease in the stress components within the cement. The results agreed with an experimental study using stems cemented into cadaver femora (96). Cement creep would accumulate for the frictional stem–cement interface, resulting in 0.46 mm total stem subsidence and a 13% decrease in the stress within the cement (92). Ling et al. (97) concluded that, at least for smooth tapered stems, that substantial hoop and radial creep of the cement not only occurs, but also is essential for the optimum clinical performance of the prosthesis. On the other hand, a limited degree of creep may help in maintaining the cement–implant interface. Also, Harris (98) stated that creep of the cement mantle surrounding a hip prosthesis may be negligible under cyclic physiological loading. The role of creep of the bone cement on the chance of failure of the cement mantle is still a subject of controversy. Due to its viscoelastic nature, cement tested at different strain rates may have changing characteristics (11). There is still lack of tests of bone cement in ways that represent real-life loading patterns that mimic those experienced by the cement *in vivo* (99). The true understanding of the mechanical behavior of bone cement can only be attained if the testing procedure is truly representative (99).

There are considerable differences between the values of physical and mechanical properties of bone cement reported in the literature. This disagreement may be the result of cement type, cement preparation technique, specimen geometry, measurement technique, test parameters, and testing environment (11). Therefore it may be impossible to compare results from different investigations (90). However, results with specific conditions may be found in a number of the excellent comprehensively reviews (1–3,6,11,22,23). It has been shown that cement formulations play a significant role on the cement mechanical properties (Table 5). Also, the test methods will affect the values obtained, which can be easily seen by comparing the bending strength results using ISO 5833 standard to that using DIN 53435 standard (Table 5). Cement mixing methods (hand mixing, vacuum mixing, or centrifugation) have been shown to affect the physical properties of bone cement (74,100,101). Even with vacuum mixing, using different vacuum mixing systems have resulted in different bone cement porosity, static strength, and fatigue strength (75,102). The storage temperature of cement constitutive was shown to have minor effects on porosity and fatigue performance (100). Ishihara et al. (103) showed that the fatigue of bone cements at 1 Hz are shorter by one to two order of magnitude as compared with fatigue lives at 20 Hz. On the other hand, Lewis et al. (104) found that frequency (over the range used) did not exert a statistically significant effect on the fatigue life of cement tested in their investigations.

Table 7. Comparison of Fatigue Test Results of Six Different Bone Cements[a]

	Cycles to Failure	
Cement Brands	Range	Weibull Median
CMW1	3042–8835	4407
CMW3	5996–38262	16441
Palacos	18362–49285	27892
Simplex P	8933–93345	36677
Osteobond	5527–25825	16162
Zimmer dough	153–3978	781

[a]See Ref. 90.

SUMMARY

Acrylic bone cement has been widely used in orthopedic surgery. Currently, the cement is not without its drawbacks as discussed previously. One major research area focusing on modifications of cement formulations may lead to obtain more favorable cement mechanical properties and biological compatibilities. One example of these developments is to incorporate second phase materials (bioactive agents, reinforcement fibers, etc.) into the existing bone cement. Another approach to overcome cement weakness is to manipulate cement processing procedure and the surgical techniques. A good example is to reduce cement porosity by vacuum mixing techniques. It is worth pointing out that considerable research is needed to develop techniques for accurate characterizations of cement properties, monitoring cement curing process and evaluations of potential bone cement failure.

BIBLIOGRAPHY

Cited References

1. Walenkamp G, Murray DW. Bone cements and cementing technique. Berlin Heidelberg New York: Springer-Verlag; 2001.
2. Kuhn KD. Bone cement: up-to-date comparison of physical and chemical properties of commercial materials. Berlin: Springer-Verlag; 2000.
3. Krause W, Mathis RS. Fatigue properties of acrylic bone cements - review of the literature. J Biomed Mater Res-Appl Biomate 1988;22(A1):37–53.
4. Charnley J. Anchorage of the femoral head prosthesis to the shaft of the femur. J Bone Joint Surg 1960;43B:28–30.

5. Kenny SM, Buggy M. Bone cements and fillers: A review. J Mater Sci-Mater Med 2003;14(11): p 923–938.
6. Lewis G. Properties of acrylic bone cement: State of the art review. J Biomed Mater Res 1997;38(2):155–182.
7. NIH. Total hip replacement. NIH Consensus Statement 1994;12(5):1–31.
8. Cristofolini L. A critical analysis of stress shielding evaluation of hip prostheses. Crit Rev Biomed Eng 1997;25(4–5):409–483.
9. Chao E. Orthopaedic Biomechanics. Int Orthop (SICOT) 1996;20:239–243.
10. Huiskes R, Verdonschot N. Biomech Art Joints: the Hip, Basic Orthopaedic Biomechanics. In: Mow VC, Hayes WC, editors. Philadelphia: Lippincott-Raven Publishers; 1997. 395–460.
11. Krause WR. Bone Cement, Acrylic. In: Webster JG, editor. Encyclopedia of Medical Devices and Instrumentation. New York: John Wiley & sons Inc.; 1988. p 491–500.
12. Wilcox RK. The biomechanics of vertebroplasty: a review. Proceedings of the Institution of Mechanical Engineers Part H. J Eng Med 2004;218(H1):1–10.
13. Phillips FM. Minimally invasive treatments of osteoporotic vertebral compression fractures. Spine 2003;28(15):S45–S53.
14. Heini PF, et al. Femoroplasty-augmentation of mechanical properties in the osteoporotic proximal femur: a biomechanical investigation of PMMA reinforcement in cadaver bones. Clin Biomech 2004;19(5):506–512.
15. Hasenwinkel J. Bone Cement. In: Wnek GE, Bowin GL, editors. Encyclopedia of Biomaterials and Biomedical Engineering. New York: Marcel Dekker; 2004. p 170–179.
16. Serbetci K, Hasirci N. Recent developments in bone cements. In: Yaszemski MJ et al. editors. Biomaterials in Orthopedics. New York: Marcel-Dekker; 2004. p 241–286.
17. Hendriks JGE, et al. Backgrounds of antibiotic-loaded bone cement and prosthesis-related infection. Biomaterials 2004;25(3):545–556.
18. Murray DW, Carr AJ, Bulstrode CJ. Which Primary Total Hip-Replacement. J Bone Joint Surg-Br Vol 1995;77B(4): 520–527.
19. Nafei A, et al. Survivorship analysis of cemented total condylar knee arthroplasty: a long-term follow-up report on 348 cases. J Arthroplasty 1996;11:7–10.
20. El-Warrak AO, et al. A review of aseptic loosening in total hip arthroplasty. Veterin Compar Orthopae Traumatol 2001;14(3):115–124.
21. Barrack RL. Early failure of modern cemented stems. J Arthroplasty 2000;15(8):1036–1050.
22. Lewis G. Fatigue testing and performance of acrylic bone-cement materials: State-of-the-art review. J Biomed Mater Res Part B-Appl Biomater 2003;66B(1):457–486.
23. Saha S, Pal S. Mechanical-Properties of Bone-Cement-a Review. J Biomed Mater Res 1984;18(4):435–462.
24. Deb S. A review of improvements in acrylic bone cements. J Biomater Appl 1999;14(1):16–47.
25. Passuti N, Gouin F. Antibiotic-loaded bone cement in orthopedic surgery. Joint Bone Spine 2003;70(3):169–174.
26. Jasty M, et al. The Initiation of Failure in Cemented Femoral Components of Hip Arthroplasties. J Bone Joint Surg Br Vol 1991;73(4):551–558.
27. Hertzberg RW, Manson JA. Fatigue of engineering plastics. London: Academic Press; 1980.
28. Spector M. Biomaterial failure. Orthoped Clin N Am 1992;23(2):211–217.
29. Pourdeyhimi B, Wagner HD. Elastic and Ultimate Properties of Acrylic Bone-Cement Reinforced with Ultra-High-Molecular-Weight Polyethylene Fibers. J Biomed Mater Res 1989;23(1):63–80.
30. Harper EJ, Behiri JC, Bonfield W. Flexural and fatigue properties of a bone cement based upon polyethylmethacrylate and hydroxyapatite. J Mat Sci Mat Med 1995;6:799–803.
31. Gilbert JL, Net SS, Lauthenschlager EP. Self-reinforced composite poly(methylmethacryate): static and fatigue properties. Biomaterials 1995;16:1043–1055.
32. Pourdeyhimi B, Wagner HD, Schwartz P. A Comparison of Mechanical-Properties of Discontinuous Kevlar-29 Fiber Reinforced Bone and Dental Cements. J Mater Sci 1986;21(12):4468–4474.
33. Wright TM, Trent PS. Mechanical-Properties of Aramid Fiber-Reinforced Acrylic Bone Cement. J Mater Sci 1979;14(2):503–505.
34. Saha S. Strain-rate dependence of the compressive properties of normal and carbon-fiber-reinforced bone-cement. J Biomed Mater Res 1983;17(6):1041–1047.
35. Saha S, et al. Biomechanical Evaluation of Bony Defects Repaired with Normal, Carbon-Fiber, and Wire Reinforced Bone-Cement. Biomater Med Devices Artif Organs 1981;9(4):291–291.
36. Pilliar RM, et al. Carbon Fiber-Reinforced Bone Cement in Orthopedic Surgery. J Biomed Mater Res 1976;10(6):893–906.
37. Topoleski LDT, Ducheyne P, Cuckler JM. The Fracture-Toughness of Titanium-Fiber-Reinforced Bone-Cement. J Biomed Mater Res 1992;26(12):1599–1617.
38. Saha S, Kraay MJ. Bending Properties of Wire-Reinforced Bone-Cement for Applications in Spinal Fixation. J Biomed Mater Res 1979;13(3):443–457.
39. Kotha SP, et al. Fracture toughness of steel-fiber-reinforced bone cement. J Biomed Mater Res Part A 2004;70A(3):514–521.
40. Fishbane BM, Pond RB. Stainless steel fiber reinforcement of polymethylmethacrylate. Clin Orthop Rel Res 1977;128:194–199.
41. Mulroy WF, Harris WH. Revision total hip arthroplasty with use of so-called second-generation cementing techniques for aseptic loosening of the femoral component—A fifteen-year-average follow-up study. J Bone Joint Surg—Am Vol 1996;78A(3):325–330.
42. Mulroy WF, Estok DM, Harris WH. Total hip arthroplasty with use of so-called second-generation cementing techniques—A fifteen-year-average follow-up study. J Bone Joint—Am Vol 1995;77A(12):1845–1852.
43. Faulkner A, et al. Effectiveness of hip prostheses in primary total hip replacement: a critical review of evidence and an economic model. Health Technol Asess 1998;2(6):1–146.
44. ASTM, ASTM Standard Specification for Acrylic Bone Cement F451-99a; 1999.
45. Dunne NJ, Orr JF. Curing characteristics of acrylic bone cement. J Mater Sci Mater Med 2002;13(1):17–22.
46. Li CD, Schmid S, Mason J. Effects of pre-cooling and preheating procedures on cement polymerization and thermal osteonecrosis in cemented hip replacements. Med Eng Phys 2003;25(7):559–564.
47. Iessaka K, Jaffe WL, Kummer FL. Effects of the initial temperature of acrylic bone cement liquid monomer on the properties of the stem-cement interface and cement polymerization. J Biomed Mater Res: Appl Biomater 2003;68B:186–190.
48. Meyer PR, Lautenschlager EP, Moore BK. On the setting properties of acrylic bone cement. J Bone Joint Surg 1973;55A:139–156.
49. Sih GC, Connelly GM, Berman AT. The Effect of Thickness and Pressure on the Curing of Pmma Bone-Cement for the Total Hip-Joint Replacement. J Biomech 1980;13(4): 347–352.

50. Vallo CL. Theoretical prediction and experimental determination of the effect of mold characteristics on temperature and monomer conversion fraction profiles during polymerization of a PMMA-based bone cement. J Biomed Mater Res (Appl Biomater) 2002;63:627–642.
51. Li CD, Mason J, Yakimicki D. Thermal characterization of PMMA-based bone cement curing. J Mater Sci—Mater Med 2004;15(1):85–89.
52. Huiskes R. Some fundamental aspects of human joint replacement. Acta Orthopaed Scand 1980; (Suppl.) 185.
53. Mjoberg B. Fixation and Loosening of Hip Prostheses—a Review. Acta Orthopaed Scand 1991;62(5):500–508.
54. Mjoberg B, et al. Bone-Cement, Thermal-Injury and the Radiolucent Zone. Acta Orthopaed Scand 1984;55(6):597–600.
55. Dipisa JA, Sih GS, Berman AT. Temperature Problem at Bone-Acrylic Cement Interface of Total Hip-Replacement. Clin Orthopae Relat Res 1976;121:95–98.
56. Swenson LW, Schurman DJ, Piziali RL. Finite element temperature analysis of a total hip replacement and measurement of PMMA curing temperatures. J Biomed Mater Res 1981;15:83–96.
57. Toksviglarsen S, Franzen H, Ryd L. Cement Interface Temperature in Hip-Arthroplasty. Acta Orthopaed Scand 1991;62(2):102–105.
58. Wang JS, Franzen H, Toksviglarsen S. Does vacuum mixing of bone-cement affect heat-generation - analysis of 4 cement brands. J Appl Biomater 1995;6(2):105–108.
59. Belkoff SM, Molloy S. Temperature measurement during polymerization of polymethylmethacrylate cement used for vertebroplasty. Spine 2003;28(14):1555–1559.
60. Moritz AR, Henriques FC. The relative importance of time and surface temperature in the causation of cutaneous burns. Am J Pathol 1947;23:695–720.
61. Lundskog J. Heat and bone tissue: an experimental investigation of the thermal properties of bone and threshold levels for thermal injury. Scand J Plastic Reconstr Surg 1972;9:1–80.
62. Revie I, Wallace M, Orr J. The effect of PMMA thickness on thermal bone necrosis around acetabular sockets. Proc Instn Mech Eng 1994;208:45–51.
63. Nelson C, Krishnan E, Neff J. Consideration of physical parameters to predict thermal necrosis in acrylic cement implants at the site of giant cell tumors of bone. Med Phys 1986;13(4):462–488.
64. Eriksson AR, Albreksson T. Temperature threshold levels for heat induced bone tissue injury: a vital microscopic study in the rabbit. J Prosthetic Den 1983;50(1):101–107.
65. Wykman AGM. Acetabular Cement Temperature in Arthroplasty - Effect of Water Cooling in 19 Cases. Acta Orthopaed Scand 1992;63(5):543–544.
66. Jasty M, et al. Porosity of various preparations of acrylic bone cements. Clin Orthop Rel Res 1990;259:122–129.
67. Muller SD, Green SM, McCaskie AW. The dynamic volume changes of polymerising polymethy methacrylate bone cement. Acta Orthopaed Scand 2002;73(6):684–687.
68. Gilbert JL, et al. A theoretical and experimental analysis of polymerization shrinkage of bone cement: A potential major source of porosity. J Biomed Mater Res 2000;52(1): 210–218.
69. Bishop NE, Ferguson S, Tepic S. Porosity reduction in bone cement at the cement-stem interface. J Bone Joint Surg-Br Vol 1996;78B:349–356.
70. James SP, et al. A fractographic investigation of PMMA bone cement focusing on the relationship between porosity reduction and increased fatigue life. J Biomed Mater Res 1992;26(5):651–662.
71. James SP, et al. Extensive porosity at the cement-femoral prosthesis interface: A preliminary study. J Biomed Mater Res 1993;27:71–78.
72. Wixson RL, Lautenschlager EP, Novak MA. Vacuum mixing of acrylic bone cement. J Arthroplasty 1987;2:141–149.
73. Topoleski LDT, Ducheyne PI, Cuckler JM. Microstructural pathway of fracture in poly(methyl methacrylate) bone cement. Biomaterials 1993;14(15):1165–1172.
74. Dunne NJ, Orr JF. Influence of mixing techniques on the physical properties of acrylic bone cement. Biomaterials 2001;22(13):1819–1826.
75. Mau H, et al. Comparison of various vaccum mixing system and bone cements as regards reliability, porosity and bending stength. Acta Orthopaed Scand 2004;75(2):160–172.
76. Iessaka K, Jaffe WL, Kummer FJ. Effects of preheating of hip prosthesis on the stem-cement interface. J Bone Joint Surg—Am Vol 2003;85:421–427.
77. Li CD, Wang Y, Mason J. The effects of curing history on residual stresses in bone cement during hip arthroplasty. J Biomed Mater Res Part B-App Biomater 2004;70B(1):30–36.
78. Race A, et al. Early cement damage around a femoral stem is concentrated at the cement/bone interface. J Biomech 2003;36:489–496.
79. Davidson CL, Feilzer AJ. Polymerization shrinkage and polymerization shrinkage stress in polymer-based restoratives. J Den 1997;25(6):435–440.
80. Davies JP, Harris WH. Comparison of diametral shrinkage of centrifuged and uncentrifuged Simplex P bone cement. J Appl Biomater 1995;6:209–211.
81. Hass S, Brauer G, Dickson G. A characterization of polymethyl methacrylate bone cement. J Bone Joint Surg—Am Vol 1975;57:380–391.
82. Ahmed AM, et al. Transient and residual stresses and displacements in self-curing bone cement - part I: characterization of relevant volumetric behavior of bone cement. J Biomech Eng—Trans ASME 1982;104:21–27.
83. Nuno N, Amabili M. Modeling debonded stem-cement interface for hip implants: effect of residual stresses. Clin Biomech 2002;17:41–48.
84. Nuno N, Avanzolini G. Residual stresses at the stem-cement interface of an idealized cemented hip stem. J Biomech 2002;35:849–852.
85. Orr JF, Dunne NJ, Quinn JC. Shrinkage stresses in bone cement. Biomaterials 2003;24(17):2933–2940.
86. Ahmed AM, et al. Transient and residual stresses and displacements in self-curing bone cement-Part II: thermoelastic analysis of the stem fixation system. J Biomech Eng—Trans Asme 1982;104:28–37.
87. Roques A, et al. Quantitative measurement of the stresses induced during polymerization of bone cement. Biomaterials 2004;25:4415–4424.
88. Zor M, Kucuk M, Aksoy S. Residual stress effects on fracture energies of cement-bone and cement-implant interfaces. Biomaterials 2002;23:1595–1601.
89. Lennon AB, Prendergast PJ. Residual stress due to curing can initiate damage in porous bone cement: experimental and theoretical evidence. J Biomech 2002;35:311–321.
90. Harper EJ, Bonfield W. Tensile characteristics of ten commercial acrylic bone cements. J Biomed Mater Res 2000;53(5):605–616.
91. Malchau H, Herberts P. Prognosis of total hip replacement and revision rate in THR: A revision risk study of 148,359 primary operations. 65th Ann Meet Am Acad Ortho Surg 1998. New Orleans.
92. Lu Z, Mckellop H. Effects of cement creep on stem subsidence and stress in the cement mantle of a total hip replacement. J Biomed Mater Res 1997;34:221–226.

93. Verdonschot N, Huiskes R. Creep properties of three low temperature-curing bone cements: a preclinical assessment. J Biomed Mater Res 2000;53B:498–504.
94. Fowler GA, et al. Experience with the Exeter total hip replacement since 1970. Ortho Clin N Am 1988;19:477–489.
95. Morgan RL, et al. Creep behavior of bone cement: a method for time extrapolation using time-temperature equivalence. J Mater Sci: Mater Med 2003;14:321–325.
96. Weightman B, et al. The mechanical properties fo cement and loosening of the femoral component of hip replacements. J Bone Joint Surg Br 1987;69:558–564.
97. Ling RSM. The use of a collar and precoating on cemented femoral stems is unnecessary and detrimental. Clin Orthop Rel Res 1992;285:73–83.
98. Harris WH. Is it advantageous to strengthen the cement-matel interface and use a collar for cemented femoral components of total hip replacement. Clin Orthop Rel Res 1992;285:67–72.
99. Eden OR, Lee AJC, Hooper RM. Stress relaxation modeling of polymethylmethacrylate bone cement. Proc Instn Mech Eng 2002;216H:195–199.
100. Lewis G. Effect of mixing method and storage temperature of cement constituents on the fatigue and porosity of acrylic bone cement. J Biomed Mater Res 1999;48B:143–149.
101. Macaulay W, et al. Difference in bone-cement porosity by vacuum mixing, centrifugation, and hand mixing. J Arthroplasty 2002;17(5):569–575.
102. Dunne NJ, et al. The relationship between porosity and fatigue characteristics of bone cements. Biomaterials 2003;24(2):239–245.
103. Ishihara S, et al. On fatigue lifetimes and fatigue crack growth behavior of bone cement. J Mater Sci Mat Med 2000;11(10):661–666.
104. Lewis G, Janna S, Carroll M. Effect of test frequency on the in vitro fatigue life of acrylic bone cement. Biomaterials 2002;24:1111–1117.

See also BIOMATERIALS, TESTING AND STRUCTURAL PROPERTIES; HIP JOINTS, ARTIFICIAL; ORTHOPEDICS, PROSTHESIS FIXATION FOR; RESIN-BASED COMPOSITES.

BONE DENSITY MEASUREMENT

YIXIAN QIN
ERIK MITTRA
Stony Brook University
New York

INTRODUCTION

Chronic diseases, such as musculoskeletal complications, have a long-term debilitating effect that greatly impacts quality of life. Osteoporosis is a reduction in bone mass or density that leads to deteriorated and fragile bones and is the leading cause of bone fractures in postmenopausal women and in the elderly population for both men and women. About 13–18% of women aged 50 years and older, and 3–6% of men aged 50 years and older, have osteoporosis in the United States alone. These rates correspond to $4 \sim 6$ million women and $1 \sim 2$ million men who suffer from osteoporosis (1). One-third of women over 65 will have vertebral fractures and 90% of women aged 75 and older have radiographic evidence of osteoporosis (2–4). Another 37–50% of women aged 50 years and older, and 28–47% of men of the same age group, have some degree of osteopenia. Thus, approximately a total of 24 million people suffer from osteoporosis in the United States alone, with an estimated annual direct cost of over $18 billion to national health programs. Hence, early diagnosis that can predict fracture risk and result in prompt treatment is extremely important. Early identification of fracture risk, most commonly caused by osteoporosis-induced bone fragility, is also important in implementing appropriate treatment and preventive strategies. Indeed, the ability to accurately assess bone fracture risk noninvasively is essential for improving the diagnostic as well as therapeutic goals (i.e., assessing temporal changes in bone during therapy) for bone loss from such varied etiologies as osteoporosis, microgravity, bed rest, or stress-shielding around an implant.

Assessment of bone mineral density (BMD) has become as essential element in the evaluation of patients at risk for osteopenia and osteoporosis (2,5–8). Bone density was initially estimated from the conventional X ray by comparing the image density of the skeleton to the surrounding soft tissues. Although demineralized bone has an image density closer to soft tissues, dense mineralized skeletal tissues appears relatively white on an X-ray image. Hence, the mineral density of bone can be estimated by the degree of gray color of the X-ray image in the bone region. However, because of its resolution and variations generated in the X-ray image, it has been suggested that bone mineral losses of at least 30% are required before they may be visually measured using a conventional X ray (9,10). Growing awareness of the impact of osteoporosis on the elderly population and the consequent costs of health care, together with the development of new treatments to prevent fractures, have led to a rapid increase in the demand for bone densitometry measurements. Many image modalities and techniques have been developed to improve the quality and the accuracy of the measurement for bone mineral and dense assessment. Two major densitometry techniques are commonly used in assessing bone density, that is, radiography-based densitometry and ultrasound-based assessment.

RADIOGRAPHY-BASED DENSITOMETRY

To improve the sensitivities of X-ray images to bone density changes and assessment, several technologies have been developed. *Bone densitometry* is a term that is defined as a method for imaging density of bone. However, the "true" density is not applicable in the current radiography-based techniques. In the field of densitometry, the term "bone mineral density", referred to as BMD, is related to the mass of bone in the tissue level, which includes both bone and marrow components as well as surrounding soft tissues. Furthermore, most densitometric techniques are projectional for the image formation that provides a two-dimensional image of the three-dimensional (3D) bone volume being measured. Therefore, the BMD defined from the projectional techniques is the mass of bone tissue mass (including marrow and/or soft surroundings) per unit area

Table 1. Radiography-Based Bone Densitometry[a]

Technique	ROI	Unit	Precision, %CV	Effective Dose, μSv
SXA	Total body	BMD ($g \cdot cm^{-2}$)	1	3
QCT	Spine	BMD ($g \cdot cm^{-2}$)	3	50–500
pQCT	Forearm	BMD ($g \cdot cm^{-2}$)	1–2	1–3
RA	Phalanx	BMD ($g \cdot cm^{-2}$)	1–2	10
SPA	Forearm	BMD ($g \cdot cm^{-2}$)	3–4	1–10
DPA	Total body	BMD ($g \cdot cm^{-2}$)	1	1–10
DXA	PA spine	BMD ($g \cdot cm^{-2}$)	1	1–10
	Proximal femur		1–2	1–10
	Total body		1	3

[a]See Refs. 5,6, 12–14.

in the image, not per unit volume of the tissue. Hence, what is actually measured is the apparent bone mineral density, which is defined by the bone mineral content contained in the area scanned, or expressed as gram per squared centimeter in unit. To detect osteoporosis accurately, several methods are developed for the noninvasive measurement of the skeleton for the diagnosis of osteopenia, osteoporosis, and/or the evaluation of an increased risk of fracture (11). These methods include single-energy X-ray absorptiometry (SXA), dual energy X-ray absorptiometry (DXA or DEXA), quantitative computed tomography (QCT), peripheral quantitative computed tomography (pQCT), radiographic absorptiometry (RA), dual photon absorptiometry (DPA), and single photon absorptiometry (SPA). There are two types of BMD measurements, peripheral BMD and central BMD. The peripheral BMD instruments are usually smaller, less expensive, and more portable than the central BMD. Central BMD is capable of measuring multiple skeletal sites, that is, the spine, the hip, and the forearm. Table 1 lists these methods currently available for the noninvasive measurement of the skeleton for the diagnosis of osteoporosis. These techniques differ substantially in physical principles, in the particular physical body sites (e.g., spine, hip, or total body), in the clinical discrimination and interpretation, and in availability of the facility and cost.

Single-Energy Densitometry

This instrumentation passes a beam of radiation through the limb of the body (e.g., forearm) and determines the difference between the incoming (or incident) radiation and the outgoing (or transmitted) radiation, referring to the attenuation. The higher the bone mineral content, the greater the attenuation. Mineral content can be calculated by the attenuation of the radiation. BMD can then be calculated by dividing the mineral content by the detected bone area. The relation between incoming and outgoing X-ray energy can be expressed as

$$I = I_0 \exp(-\lambda d) \tag{1}$$

where I_0 = incoming radiation intensity, I = transmitted radiation intensity, λ = mass attenuation coefficient, and d = area density of the attenuating materials ($g \cdot cm^{-2}$). The mass attenuation coefficient is a physical property that describes how much a given material attenuates and X-ray energy. If the attenuation coefficient can be experimentally determined, the equation becomes explicit, and the area density can be determined by

$$d = k \log(I_0/I) \tag{2}$$

where k is an experimentally determined constant for the attenuation coefficient.This technology is relatively simple and easy to understand. However, biological tissues and body are composed by multiple materials (e.g., bone, muscle, and other soft tissues). The accuracy of the technology is limited.

Single-Photon Absorptiometry

Bone density can be measured by passing a monochromatic or single-energy photon beam through bone and soft tissue. This procedure is referred as SPA. The amount of mineral content can be quantified by the attenuation of the beam intensity. After the photon beam attenuation is calculated, the value of the attenuation can be compared with a calibration parameter derived from a standard mineral content (e.g., using ashed bone of known weight). This procedure can finally determine the BMD with measured attenuation. Iodine-125 at 27 keV, or americium-241 at 59.5 keV, was initially used for generating of the SPA beam. SPA is rarely used in clinical practice today. SPA determined bone mineral content is calculated through uniform thickness of the soft tissue in the path of the beam. The targeted scan site (e.g., limb or forearm) had to be submerged in the water or a tissue-equivalent material, which limited the practical applications of the SPA. The advantages of this technique include a low dose of radiation, portable, and use for particular body sites with relatively precisely measurement. Although the SPA is an approximate method, the limitations of SPA include limited accuracy of the measurement, radiation, and used only on the particular peripheral sites, like forearm and heel.

Dual-Photon Absorptiometry

To overcome the limitations of single-energy or photon densitometry, if a dual radiation source was used, the influence of soft tissues could be eliminated. The basic principle involved in DPA for bone density measurement was similar to SPA. The degree of attenuation of the photon energy beam between incoming and outgoing energy through bone and soft tissue is quantified. As with SPA, the beam source was originally used, but with an isotope

used, which emitted photon energy at two distinct photoelectric peaks. When the beam was passed through a region of the body with both hard and soft tissues, attenuation of the photon beam appeared to both photon energy peaks. The contributions of soft tissue to beam attenuation can be determined by the quantifications of the relative relations between two attenuations (15). Because of its capability to distinct bone from soft tissue, DPA has been used to quantify bone density in deep tissue and large skeletal areas where bone is surrounded by large volume of soft masses (e.g., spine and hip) (16). DPA was considered a major advance from SPA due to its ability to quantify BMD and mineral content in such deep areas like spine and hip, as well as its capability of quantifications of effects from soft tissues. However, DPA has many notable limitations. First, the maintenance of the beam source was expensive, which had to be replaced yearly. Second, the radioactive source decay increased as much as 0.6% per month, which added difficulties for the calibration. These factors may result in the precision of 2–4% for DPA measurements in the region of interests. This precision (e.g., 2%) would limit its clinical application, in which a great change (e.g., 5–6%) from the baseline value had to be observed before one could reach the 95% confidence level for the change of bone density. Nevertheless, the concept of dual-photon densitometry has impacted the development of new technologies such as DXA.

Dual-Energy X-ray Absorptiometry

Perhaps the most popular bone densitometry used in clinical practice is the DXA or DEXA. The basic principles of DXA are the same as DPA. To overcome the major limitation of DPA, it did not take long for manufacturers who originally had the DPA product to replace the decaying isotope beam source with a highly stable dual-energy X-ray tube. There are several advantages of using X-ray sources over radioactive isotopes (i.e., no beam decay concerned in the X-ray tube and no calibration required for correction of the drifting because of the source decay in the DPA). The fundamental basis for DXA is the measurement of the transmission through the body of X rays of two different photon energies. The radiation source is collimated to a pencil beam and aimed at a radiation detector placed directly opposite the objective to be measured (Fig. 1). The patients are positioned on a table in the path of the X-ray beams. Due to the dependence of the attenuation coefficient on atomic number and photon energy, assessment of the transmission factors and attenuations at two energies enables the 2D apparent density, that is, bone density per unit projected area, of two different types of tissues to be inferred (17–19). The X-ray source and detector pair is scanned back and forth across the region of interests in the body, generating annotation images, which the BMD is calculated as the ratio of the bone content to the measured area. Radiation dose to the patients is very low on the order of 1–10 μSv. The DXA system can measure the BMD of the spine, proximal femur, forearm, and the total body. Recent technology uses a fan beam geometry in the DXA scanners that can increase the speed and reduce the acquisition time (GE Medical System Inc.). The image

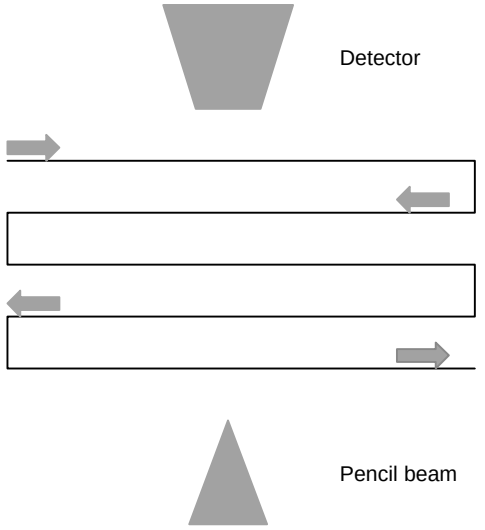


Figure 1. Scan path pattern for DXA densitometer using pencil-beam format.

quality of recent DXA has improved significantly via computational capability for better visualization. In addition to the body DXA scanners, recent technology has also adapted for development of lower cost, small, and particular site densitometries (Fig. 2). These systems are available to the clinic for specific body regions (e.g., spine, hip, leg, arm, and hand). The DXA systems are available for the diagnostic clinical use by many major manufacturers (e.g., GE Medical Systems of Madison, Norland, and Hologic Inc. of Bedford).

The basic working principle of DXA and its ability to reduce the effects of soft tissue is to use two X-ray sources and mathematically solve the bone thickness and soft-tissue thickness (15). By using two X-ray energies, two equations can be derived by scanning the measurement site twice with low (L) and high (H) energies once at each.

$$I^L = I_0^L[\exp - (\lambda_b^L d_b + \lambda_s^L d_s)] \quad (3)$$

$$I^H = I_0^H[\exp - (\lambda_b^H d_b + \lambda_s^H d_s)] \quad (4)$$

where $I_0 =$ incoming radiation intensity, $I =$ transmitted radiation intensity, $\lambda =$ mass attenuation coefficient, $d =$ area density of the attenuating materials ($g \cdot cm^{-2}$), and b and s refer to the bone and soft tissue. Two scans are usually performed simultaneously with either two energies or rapid switching between two energies. When the

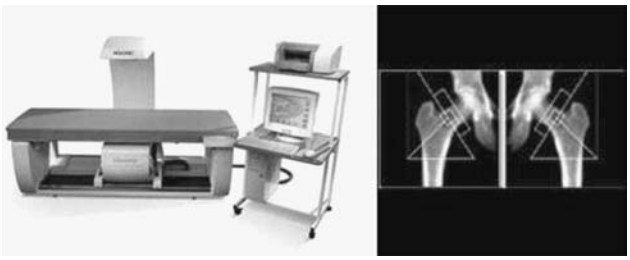


Figure 2. DEXA machine for a whole-body scan (QDR4500 fan-beam scanner, Hologic Inc., Bedford, MA) (left). DEXA bone densitometry is widely used in.

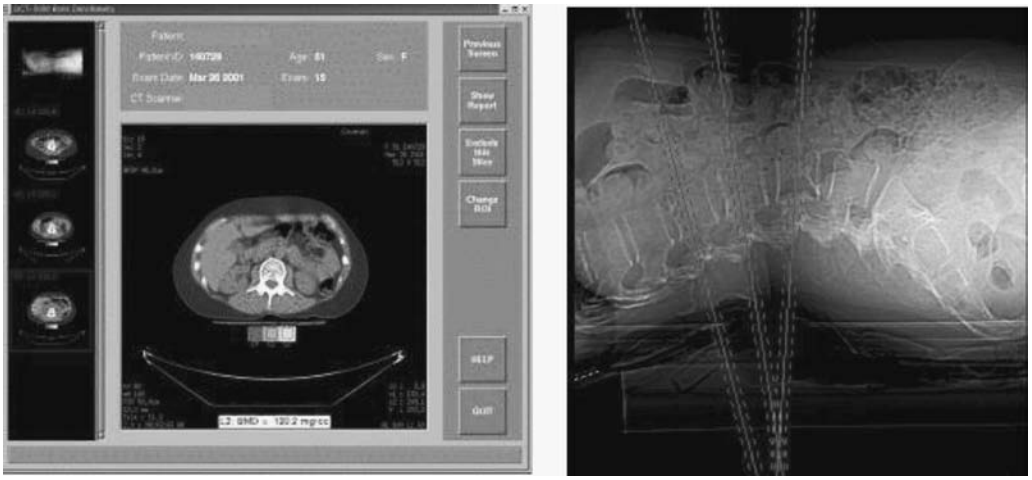


Figure 3. QCT allows selection of the region of interest.

attenuation coefficients for bone and soft tissue are known for both low and high energies, the apparent or area bone mineral density can be calculated as

$$d_b = \frac{(\lambda_s^L/\lambda_s^H)\log(I^H/I_0^H) - \log(I^L/I_0^L)}{\lambda_b^L - \lambda_b^H(\lambda_s^L/\lambda_s^H)} \quad (5)$$

The attenuation coefficient for bone is relative constant but varied to persons. The soft-tissue attenuation coefficient, however, is contributed by fat and other soft tissues and varied greatly in the body. This is the source for the errors generated in the measurement. The manufacturers usually provide phantoms for calibration for the system.

Traditionally, the focus of clinical bone evaluation has been apparent or area BMD as measured by DXA or DEXA (20–23). DEXA provides an effective way to measure BMD in a specific region of interest and is the most widely used diagnostic modality for assessing osteoporosis and osteopenia (8,24–26). However, in particular, DEXA suffers from several shortcomings. Although density (quantity) does positively correlate with strength (27,28) and fracture risk (29–31), anywhere from 10–90% of the variability in bone strength remains unexplained the (32). Additionally, as discussed below, the stereology of trabecular bone is one of its distinguishing features (especially with respect to its mechanical behavior), but because DEXA provides only a 2D image of apparent density, it is inherently limited in this regard. DEXA also suffers from an inability to differentiate trabecular from cortical bone. Although it is true that cortical bone also deteriorates with age (33,34), the effects of bone loss are more prevalent in trabecular bone due to its much higher surface area, and the greater net amount of bone mineral content in cortical bone can conceal small changes in the trabecular bone when measuring only BMD. Nevertheless, as one of the key factors that contribute to the bone's quality evaluation, BMD measured by DEXA is a most popular modality used in assessing the status of bone and the risk of fracture.

Quantitative Computed Tomography (QCT)

QCT provides the true volumetric 3D bone density ($mg \cdot cm^{-3}$) compared with the 2D apparent or aeral density measurement with DEXA (35–43). Because of its high resolution, QCT can provide the measurement in the trabecular region (e.g., femoral neck and vertebral bodies) (39,40,44,45). Compared with DXA the advantage of QCT is the image-based cross-sectional anatomy, which allows for a selection of the region of interests (ROI) and a better assessment of geometrical properties (Fig. 3). Most CT systems provide a software package to automate the placement of the ROI within a particular body volume, e.g., vertebral bodies. QCT scans are generally performed using a single kilovolt setting (single-energy QCT). It is possible to use a dual-energy QCT, which can provide further improvement of the resolution, but at the price of poorer precision and higher radiation dose. New 3D volumetric techniques acquire datasets with which analysis of bone macroarchitecture may be further optimized. Due to its capability of high resolution, geometric and structural parameters determined in QCT may contribute to determine bone strength when integrated with other technology (i.e., finite element analysis). The advantage of spinal QCT is the high responsiveness of the vertebral trabecular bone to aging and disease, whereas the principal disadvantage is the cost of the equipment and the dosage received for the scanning (higher than DEXA).

Peripheral QCT (pQCT)

pQCT systems are available for measuring the forearm. The advantages of these devices are the capability of separating the trabecular and cortical bone of the ultradistal radius and of reporting volumetric density. Several clinical used pQCT devices are available (e.g., the Stratec XCT 2000, that are suitable for use in a physician's office or in primary care).

QUANTITATIVE ULTRASONOMETRY

Quantitative ultrasound (QUS) for measuring the peripheral skeleton has raised considerable interest in recent years. New methods have emerged with the potential to estimate trabecular bone modulus more directly. QUS provides an intriguing method for characterizing the

Table 2. Summary of Current QUS Devices for Calcaneus

Device	Performance	Resolution	Predict Parameter	Cost, $K
Sahara (Hologic)	Index	Nonimage	Z score	20–25
QUS-2 (Metra Biosystem)	Index	Nonimage	Z score	20–25
UBA 575 (Walker Sonix)	Index	Nonimage	Z score	20–25
Achilles (GE-Lunar)	Index + image	Image, 5 mm	Z score	40–50
UBIS 5000 (DMS)	Index + image	Image, 2 mm	Z score	30–35
DTU-one (Osteometer)	Index + image	Image, 2 mm	Z score	25–30
New SCAN	Image + index	Image, 1 mm	Stiffness, BMD, Z	20

material properties of bone in a manner that is noninvasive, nonionizing, nondestructive, and relatively accurate. The primary advantage of QUS is that it is capable of measuring not only bone quantity (e.g., BMD), but also bone quality (i.e., estimation of the mechanical property) of bone. Over the past 15 years, several research approaches have been developed to quantitate bone mass and structural stiffness using QUS (46–48). Preliminary results for predicting osteoporosis using QUS are promising, and it has great potential for widespread applications (including screening for prevention). As such, many QUS machines have been developed, and there are currently many different devices on the market. Most available systems measure the calcaneus using plane waves that use either water or gel coupling [e.g., Sahara (Hologic Inc., MA), QUS-2 (Metra Biosystems Inc., CA), Paris (Norland Inc., WI), and UBA 575 (Walker Sonix Inc., USA)] (Table 2). Recently, an image-based bone densitometry device for calcaneous ultrasound measurement is also made available using an array of plane ultrasound wave (GE-Lunar, Inc., USA). Using several available clinical devices, studies in vivo have shown the ability of QUS to discriminate patients with osteoporotic fractures from age-matched controls (49–51). It has been demonstrated that QUS predicts risk of future fracture generally as well as DEXA (51–54). However, there are several noted limitations, including the tissue boundary interaction, the nonlinear function of density associated with bone ultrasonic attenuation, the single index covering a broad range of tissues (including the cortical and trabecular regions), and the interpolation of the results. Recently, a focused ultrasound sonometer device was developed to obtain the likelihood of a broadband ultrasound attenuation (BUA) image in the human calcaneus region (center frequency 0.5 MHz, focus 50 mm) (55,56) (UBIS 5000, Diagnostic Medical Systems; and DTU-one, Osteometer MediTech). These devices provide ultrasound images in the calcaneus region, in which the parameter compares with DEXA data. Perhaps the major drawbacks of these ultrasound osteometers are low resolution and lack of physical interrelation with meaningful bone strength. Although only showing the correlation between BUA data and BMD, these devices mostly provide qualitative information for assessment of osteoporosis, not the true prediction for bone structural and strength properties. Therefore, QUS remains at a stage as a screening tool (Fig. 4), because of the nonuniformity of the porous structure in the bone tissue and its associated effects in resolution (14). Research attention is focused on developing systems to provide true images reflecting the bone's structural and strength properties at multiple skeletal sites, i.e., in the hip, which can provide a true diagnostic tool (instead of just for screening) that surpasses the radiation based DEXA machines.

If QUS bone densitometry can be developed to provide a "true" bone quality parameter-based diagnostic tool (i.e., directly related to the bone's structural and strength properties) and to target multiple and critical skeletal sites (e.g., hip and distal femur), QUS would have a greater impact on the diagnosis of bone diseases (e.g., osteoporosis) than current available bone densitometry. Research efforts are made in this regard (55–59). As an example, a new QUS modality, called the scanning confocal acoustic diagnostic system, has been developed (57–60), which is intended to provide true images reflecting the bone's structural and strength properties at a particular skeletal site at a peripheral limb and potentially at deep tissue like great trachanter. The technology may further provide both density and strength assessment in the region of interests for the risk of fracture (57–60).

Fundamental QUS Parameters in Bone Measurement

In an effort to use QUS for predicting bone quality, a variety of approaches have been explored with many studies published in the past decade, that have examined the utility of QUS and its potential application as a diagnostic tool for osteoporosis. The physical mechanisms of ultrasound applied to bone may include several fundamental approaches, [i.e., speed of sound (SOS) or ultrasonic wave propagating velocity (UV), sound energy attenuation

Figure 4. A QUS bone densitometry test in a heel region. Reproduced courtesy of GE-Lunar Inc.

(ATT), BUA, and critical angle ultrasound parameters] that closely relate to acoustic transmission in a porous structure. Most commonly, parameters for QUS measurement are BUA and SOS, which can be used to identify those persons at risk of osteoporotic fracture as reliably as BMD (52–54,61,62). It has been shown that both BUA and SOS are decreased in persons with risk factors for osteoporosis, that is, primary hyperparathyroidism (63–66), kidney disease (67), and glucocorticoid use (68,69). The proportion of women classified into each diagnostic category was similar for BMD and QUS. Using the World Health Organization (WHO) criteria to classify osteoporosis for BMD measurement using DEXA and QUS testing, approximately one third of postmenopausal women aged 50+ years with clinical risk factors were diagnosed as osteoporotic compared with only 12% of women without clinical risk factors. This suggests that the measurement of QUS with calcaneal BUA and SOS is to some extent the same as the BMD Z-score measurement.

Background of BUA in Trabecular Bone Measurement

BUA and SOS are currently two commonly used methods for QUS measurements, which make it potentially possible to predict bone density and strength. As an ultrasound wave propagates through a medium, BUA measures the acoustic energy that is lost in bone (unit: dB/MHz). The slope at which attenuation increases with frequency is generally between 0.2 and 0.6 MHz, and it characterizes BUA. The slopes of the frequency spectrum may reflect the density and structure of bone. Although relatively little is known about the fundamental interactions that determine ultrasound attenuation in bone, the potential sources contributing to the attenuation include absorption, scattering, diffraction, and refraction (70–73). Although absorption predominates in cortical bone attenuation, the mechanism of BUA in cancellous bone is believed to be scattering (14,74–76). The importance of scattering has been alluded to in the literature. Scattering is also suggested to contribute to the nonlinear variation in BUA with density observed in cancellous bone and a porous medium (77–79).

Background of SOS or UV for Bone Measurement

The strength of trabecular bone is an important parameter for bone quality. *In vitro* studies have correlated the ultrasound velocity with stiffness in trabecular bone samples (80–82). This indicates that ultrasound has the potential to be advantageous over the X-ray based absorptiometry in assessing the quality of bone in addition to the quantity of bone. The mechanism of SOS in predicting bone strength is believed to be due to the fact that the velocity of an ultrasound wave depends on the material properties of the medium through which it is propagating, but it also depends on the mode of propagation. By determining the wave velocity through a bone, the elastic modulus of bone specimens can be evaluated, or at least be approximated (80,83). When ultrasound travels through a porous material, e.g., trabecular bone, it carries information concerning material properties, such as density, elasticity, and architecture. A relationship exists between the ultrasound velocity (unit: m/s) and the material elasticity E and density ρ (14,80)

$$V = \sqrt{E/\rho} \tag{6}$$

The velocity with which ultrasound passes through normal bone is fast and varies depending on whether the bone is cortical or trabecular. Speeds of 2800–3000 $\mathrm{m \cdot s^{-1}}$ are typical in cortical bone, whereas speeds of 1550–2300 $\mathrm{m \cdot s^{-1}}$ are typical in trabecular bone.

It is demonstrated that trabecular bone strength is highly correlated with elastic stiffness (84). With the introduction of QUS, several new diagnostic parameters and experimental results, both *in vitro* and *in vivo*, have shown potential for evaluating not only bone quantity (i.e., BMD), but also bone quality (i.e., structure and strength). Two principal variables, BUA and UV, have been confirmed to identify those persons at risk of osteoporotic fracture as reliably as BMD from DEXA. However, SOS and BUA are related to bone density and strength as well as to trabecular orientation, the proportion of trabecular bone and cortical shell, the composition of organic and inorganic components, and the conductivity of the cancellous structure. Thus, QUS of trabecular bone depends on a variety of factors that contribute to the measured ultrasound parameters.

Other Bone Status Measurement Methods and Motivation to Assess Bone Quality

Beyond bone quantity, the quality (the integrity of its structure and strength) has become an equally or even more important measure to understand the bone structure and mechanical integrity. Most osteoporotic fractures occur in cancellous bone. Therefore, noninvasive assessment of trabecular bone strength and stiffness is extremely important in predicting the quality of the bone. The strength of the trabecular bone mostly depends on the mechanical properties of the bone at the local and bulk tissue level, and on its spatial distribution (i.e., the microarchitecture). A better understanding of the factors that influence bone strength is a key to developing improved diagnostic techniques and more effective treatments. To overcome the current hurdles, to improve the "quality" of the noninvasive diagnostic instrumentations, and to apply the technology for future clinical application, new clinical modality may concentrate in several main areas: (*1*) increasing the resolution, sensitivity, and accuracy in diagnosing osteoporosis through unique methods for improvement of signal/noise ratio; (*2*) directly measuring bone's strength as one of the primary parameters for the risk of fracture; (*3*) generating real-time compatible imaging to identify local region of interest; (*4*) validating structural and strength properties with new modalities; and (*5*) predicting local trabecular and bulk stiffness and microstructure of bone, and generating a physical relationship between measurement and bone quality. In an attempt to achieve these goals, recent advances of emerging technologies are developed primarily for animal studies at this stage. These include high resolution pQCT, micro-MR-derived measures of structure, micro-CT-based BMD, and combined assessment of strength using geometry, density, and computational simulation. These methods

will further lead to a better understanding of the progressive deterioration of bone in aging populations, and ultimately they may provide early prediction of fracture risk and associated musculo-skeletal complications such as osteoporosis.

ACKNOWLEDGMENT

This work has been kindly supported by the National Space Biomedical Research Institute (TD00207 and TD00405 to Y. Qin) through NASA Cooperative Agreement NCC 9-58.

BIBLIOGRAPHY

Cited References

1. Looker AC, Johnson CL. Prevalence of elevated serum transferrin saturation in adults in the United States. Ann Intern Med 1998;129:940–945.
2. Melton LJI. How many women have osteoporosis now?. J Bone Min Res 1995;10:175–177.
3. Melton LJI. Epidemiology of hip fracture: Implications of the exponential increase with age. Bone 1996;18:121S–125S.
4. Wahner HW, Fogelman I. The Evaluation of Osteoporosis: Dual Energy X-Ray Absorptiometry in Clinical Practice. London: 1994.
5. Genant HK. Current state of bone densitometry for osteoporosis. Radiographics 1998;18:913–918.
6. Kanis JA. An update on the diagnosis of osteoporosis. Curr Rheumatol Rep 2000;2:62–66.
7. Melton LJ III, Atkinson EJ, O'Connor MK, O'Fallon WM, Riggs BL. Bone density and fracture risk in men. J Bone Miner Res 1998;13:1915–1923.
8. Melton LJ III, Orwoll ES, Wasnich RD. Does bone density predict fractures comparably in men and women? Osteoporos Int 2001;12:707–709.
9. Sartoris DJ, Resnick D. Current and innovative methods for noninvasive bone densitometry. Radiol Clin North Am 1990;28:257–278.
10. Sartoris DJ, Resnick D. X-ray absorptiometry in bone mineral analysis. Diagn Imaging (San Franc) 1990;12: 108–113,159,183.
11. Lewiecki EM. Clinical applications of bone density testing for osteoporosis. Minerva Med 2005;96:317–330.
12. Blake G. M, Gluer CC, Fogelman I. Bone densitometry: Current status and future prospects. Br J Radiol 1997;70: Spec No:S177–S186.
13. Blake GM, Fogelman I. Bone densitometry and the diagnosis of osteoporosis. Semin Nucl Med 2001;31:69–81.
14. Njeh CF, Hans D, Fuerst T, Gluer C-C, Genant HK. Quantitative Ultrasound Assessment of Osteoporosis and Bone Status. Munich: 1999.
15. Nord RH. Technical consideration in DPA. In: Genany HK, editor. Osteoporosis Updates 1987. 1987:203–212.
16. Dunn WL, Wahner HW, Riggs BL. Measurement of bone mineral content in human vertebrae and hip by dual photon absorptiometry. Radiology 1980;136:485–487.
17. Blake G. M, Fogelman I. Dual energy x-ray absorptiometry and its clinical applications. Semin Musculoskelet Radiol 2002;6:207–218.
18. Blake G. M, Fogelman I. Methods and clinical issues in bone densitometry and quantitatie ultrasonometry. 1573–1585, 2002.
19. Blake G. M, Fogelman I. Fracture prediction by bone density measurements at sites other than the fracture site: The contribution of BMD correlation. Calcif Tissue Int 2005;76:249–255.
20. Kanis JA. Diagnosis of osteoporosis and assessment of fracture risk. Lancet 2002;359:1929–1936.
21. Kanis JA. Assessing the risk of vertebral osteoporosis. Singapore Med J 2002;43:100–105.
22. Kanis JA, Borgstrom F, Zethraeus N, Johnell O, Oden A, Jonsson B. Intervention thresholds for osteoporosis in the UK. Bone 2005;36:22–32.
23. Kanis JA, Borgstrom F, De Laet C, Johansson H, Johnell O, Jonsson B, Oden A, Zethraeus N, Pfleger B, Khaltaev N. Assessment of fracture risk. Osteoporos Int 2005;16:581–589.
24. Melton LJ III, Atkinson EJ, O'Connor MK, O'Fallon WM, Riggs BL. Determinants of bone loss from the femoral neck in women of different ages. J Bone Miner Res 2000;15:24–31.
25. Melton LJ III, Kanis JA, Johnell O. Potential impact of osteoporosis treatment on hip fracture trends. J Bone Miner Res 2005;20:895–897.
26. Vokes TJ, Favus MJ. Noninvasive assessment of bone structure. Curr Osteoporos Rep 2003;1:20–24.
27. Keaveny TM, Morgan EF, Niebur GL, Yeh OC. Biomechanics of trabecular bone. Annu Rev Biomed Eng 2001;3:307–333.
28. Keaveny TM, Yeh OC. Architecture and trabecular bone—toward an improved understanding of the biomechanical effects of age, sex and osteoporosis. J Musculoskelet Neuronal Interact 2002;2:205–208.
29. Johnston CC Jr, Slemenda CW. Risk assessment: Theoretical considerations. Am J Med 1993;95:2S–5S.
30. Johnston CC Jr, Slemenda CW. Peak bone mass, bone loss and risk of fracture. Osteoporos Int 1994;4 (Suppl 1):43–45.
31. Johnston CC Jr, Hui S. Absolute versus relative fracture risk. J Bone Miner Res 2005;20:704.
32. Hans D, Fuerst T, Lang T, Majumdar S, Lu Y, Genant HK, Gluer C. How can we measure bone quality? Baillieres Clin Rheumatol 1997;11:495–515.
33. Dempster DW, Ferguson-Pell MW, Mellish RW, Cochran GV, Xie F, Fey C, Horbert W, Parisien M, Lindsay R. Relationships between bone structure in the iliac crest and bone structure and strength in the lumbar spine. Osteoporos Int 1993;3:90–96.
34. Dempster DW, Cosman F, Kurland ES, Zhou H, Nieves J, Woelfert L, Shane E, Plavetic K, Muller R, Bilezikian J, Lindsay R. Effects of daily treatment with parathyroid hormone on bone microarchitecture and turnover in patients with osteoporosis: A paired biopsy study. J Bone Miner Res 2001;16:1846–1853.
35. Laib A, Hauselmann HJ, Ruegsegger P. In vivo high resolution 3D-QCT of the human forearm. Technol Health Care 1998; 6:329–337.
36. Lang T, Augat P, Majumdar S, Ouyang X, Genant HK. Noninvasive assessment of bone density and structure using computed tomography and magnetic resonance. Bone 1998;22:149S–153S.
37. Lang TF, Keyak JH, Heitz MW, Augat P, Lu Y, Mathur A, Genant HK. Volumetric quantitative computed tomography of the proximal femur: Precision and relation to bone strength. Bone 1997;21:101–108.
38. Lang TF, Augat P, Lane NE, Genant HK. Trochanteric hip fracture: Strong association with spinal trabecular bone mineral density measured with quantitative CT. Radiology 1998;209:525–530.
39. Lang TF, Li J, Harris ST, Genant HK. Assessment of vertebral bone mineral density using volumetric quantitative CT. J Comput Assist Tomogr 1999;23:130–137.
40. Lang TF, Guglielmi G, Kuijk Cvan, De Serio A, Cammisa M, Genant HK. Measurement of bone mineral density at the spine and proximal femur by volumetric quantitative com-

puted tomography and dual-energy X-ray absorptiometry in elderly women with and without vertebral fractures. Bone 2002;30:247–250.
41. Ruegsegger P, Stebler B, Dambacher M. Quantitative computed tomography of bone. Mayo Clin Proc 1982;57 (Suppl):96–103.
42. Ruegsegger P. Quantitative computed tomography at peripheral measuring sites. Ann Chir Gynaecol 1988;77:204–207.
43. Ruegsegger P, Steiger P, Felder M. Quantitative computed tomography of the rheumatic knee. Clin Rheumatol 1988; 7:486–491.
44. Cann CE, Genant HK, Kolb FO, Ettinger B. Quantitative computed tomography for prediction of vertebral fracture risk. Bone 1985;6:1–7.
45. Cann CE. Quantitative CT for determination of bone mineral density: A review. Radiology 1988;166:509–522.
46. Ashman RB, Cowin SC, Van Buskirk WC, Rice JC. A continuous wave technique for the measurement of the elastic properties of cortical bone. J Biomech 1984;17:349–361.
47. Ashman RB, Corin JD, Turner CH. Elastic properties of cancellous bone: measurement by an ultrasonic technique. J Biomech 1987;20:979–986.
48. Ashman RB, Rho JY. Elastic modulus of trabecular bone material. J Biomech 1988;21:177–181.
49. Cheng S, Tylavsky F, Carbone L. Utility of ultrasound to assess risk of fracture. J Am Geriatr Soc 1997;45:1382–1394.
50. Gregg EW, Kriska AM, Salamone LM, Roberts MM, Anderson SJ, Ferrell RE, Kuller LH, Cauley JA. The epidemiology of quantitative ultrasound: A review of the relationships with bone mass, osteoporosis and fracture risk. Osteoporos Int 1997;7:89–99.
51. Njeh CF, Boivin CM, Langton CM. The role of ultrasound in the assessment of osteoporosis: A review. Osteoporos Int 1997;7:7–22.
52. Bauer DC, Gluer CC, Cauley JA, Vogt TM, Ensrud KE, Genant HK, Black DM. Broadband ultrasound attenuation predicts fractures strongly and independently of densitometry in older women. A prospective study. Study of Osteoporotic Fractures Research Group. Arch Intern Med 1997;157:629–634, 3–24.
53. Hans D, Schott AM, Meunier PJ. Ultrasonic assessment of bone: A review. Eur J Med 1993;2:157–163.
54. Hans D, Schott AM, Arlot ME, Sornay E, Delmas PD, Meunier PJ. Influence of anthropometric parameters on ultrasound measurements of Os calcis. Osteoporos Int 1995;5:371–376.
55. Laugier P, Fournier B, Berger G. Ultrasound parametric imaging of the calcaneus: *In vivo* results with a new device. Calcif Tissue Int 1996;58:326–331.
56. Laugier P, Droin P, Laval-Jeantet AM, Berger G. In vitro assessment of the relationship between acoustic properties and bone mass density of the calcaneus by comparison of ultrasound parametric imaging and quantitative computed tomography. Bone 1997;20:157–165.
57. Qin Y-X, Lin W, Rubin C. Interdependent relationship between Trabecular bone quality and ultrasound attenuation and velocity using a scanning confocol acoustic diagnostic system. J Bone Min Res 2001;16:S470–S470.
58. Qin Y-X, Lin W, Mittra E, Mueller R, Xia Y, Rubin C. Noninvasive assessment of bone quality and quantity using confocal acoustic scanning on *ex-vivo* trabeculae. Ann Biomed Eng. In press.
59. Qin Y-X, Xia Y, Lin W, Chadha A, Gruber B, Rubin C. Assessment of bone quantity and quality in human cadaver calcaneus using scanning confocal ultrasound and DEXA measurements. J Bone Min Res 2002;17:S422–S422.
60. Xia Y, Lin W, Qin Y. The influence of cortical end-plate on broadband ultrasound attenuation measurements at the human calcaneus using scanning confocal ultrasound. J Acoustic Soc Am 2005;118:1801–1807.
61. Frost ML, Blake GM, Fogelman I. Contact quantitative ultrasound: An evaluation of precision, fracture discrimination, age-related bone loss and applicability of the WHO criteria. Osteoporos Int 1999;10:441–449.
62. Frost ML, Blake GM, Fogelman I. Quantitative ultrasound and bone mineral density are equally strongly associated with risk factors for osteoporosis. J Bone Miner Res 2001;16:406–416.
63. Gomez AC, Schott AM, Hans D, Niepomniszcze H, Mautalen CA, Meunier PJ. Hyperthyroidism influences ultrasound bone measurement on the Os calcis. Osteoporos Int 1998;8:455–459.
64. Guo CY, Thomas WE, al Dehaimi AW, Assiri AM, Eastell R. Longitudinal changes in bone mineral density and bone turnover in postmenopausal women with primary hyperparathyroidism. J Clin Endocrinol Metab 1996;81:3487–3491.
65. Minisola S, Scarnecchia L, Carnevale V, Bigi F, Romagnoli E, Pacitti MT, Rosso R, Mazzuoli GF. Clinical value of the measurement of bone remodelling markers in primary hyperparathyroidism. J Endocrinol Invest 1989;12:537–542.
66. Minisola S, Rosso R, Scarda A, Pacitti MT, Romagnoli E, Mazzuoli G. Quantitative ultrasound assessment of bone in patients with primary hyperparathyroidism. Calcif Tissue Int 1995;56:526–528.
67. Wittich A, Vega E, Casco C, Marini A, Forlano C, Segovia F, Nadal M, Mautalen C. Ultrasound velocity of the tibia in patients on haemodialysis. J Clin Densitometry 1998;1:157–163.
68. Blanckaert F, Cortet B, Coquerelle P, Flipo RM, Duquesnoy B, Marchandise X, Delcambre B. Contribution of calcaneal ultrasonic assessment to the evaluation of postmenopausal and glucocorticoid-induced osteoporosis. Rev Rhum Engl Ed 1997;64:305–313.
69. Cortet B, Flipo RM, Blanckaert F, Duquesnoy B, Marchandise X, Delcambre B. Evaluation of bone mineral density in patients with rheumatoid arthritis. Influence of disease activity and glucocorticoid therapy. Rev Rhum Engl Educ 1997;64:451–458.
70. Madsen EL, Dong F, Frank GR, Garra BS, Wear KA, Wilson T, Zagzebski JA, Miller HL, Shung KK, Wang SH, Feleppa EJ, Liu T, O'Brien WD Jr, Topp KA, Sanghvi NT, Zaitsev AV, Hall TJ, Fowlkes JB, Kripfgans OD, Miller JG. Interlaboratory comparison of ultrasonic backscatter, attenuation, speed measurements. J Ultrasound Med 1999;18:615–631.
71. Wear KA, Garra BS. Assessment of bone density using ultrasonic backscatter. Ultrasound Med Biol 1998;24: 689–695.
72. Wear KA. Frequency dependence of ultrasonic backscatter from human trabecular bone: theory and experiment. J Acoust Soc Am 1999;106:3659–3664.
73. Wear KA, Stuber AP, Reynolds JC. Relationships of ultrasonic backscatter with ultrasonic attenuation, sound speed and bone mineral density in human calcaneus. Ultrasound Med Biol 2000;26:1311–1316.
74. Strelitzki R, Evans JA. An investigation of the measurement of broadband ultrasonic attenuation in trabecular bone. Ultrasonics 1996;34:785–791.
75. Strelitzki R, Evans JA. Diffraction and interface losses in broadband ultrasound attenuation measurements of the calcaneum. Physiol Meas 1998;19:197–204.
76. Strelitzki R, Metcalfe SC, Nicholson PH, Evans JA, Paech V. On the ultrasonic attenuation and its frequency dependence

in the os calcis assessed with a multielement receiver. Ultrasound Med Biol 1999;25:133–141.
77. Aindow JD, Chivers RC. Ultrasonic wave fluctuations through tissue: An experimental pilot study. Ultrasonics 1988;26:90–101.
78. Chivers RC. The scattering of ultrasound by human tissues–some theoretical models. Ultrasound Med Biol 1977;3:1–13.
79. Chivers RC, Parry RJ. Ultrasonic velocity and attenuation in mammalian tissues. J Acoust Soc Am 1978;63:940–953.
80. Ashman RB, Rho JY, Turner CH. Anatomical variation of orthotropic elastic moduli of the proximal human tibia. J Biomech 1989;22:895–900.
81. McKelvie ML, Palmer SB. The interaction of ultrasound with cancellous bone. Phys Med Biol 1991;36:1331–1340.
82. Turner CH, Eich M. Ultrasonic velocity as a predictor of strength in bovine cancellous bone. Calcif Tissue Int 1991;49:116–119.
83. Rho JY, Ashman RB, Turner CH. Young's modulus of trabecular and cortical bone material: Ultrasonic and microtensile measurements. J Biomech 1993;26:111–119.
84. Hou FJ, Lang SM, Hoshaw SJ, Reimann DA, Fyhrie DP. Human vertebral body apparent and hard tissue stiffness. J Biomech 1998;31:1009–1015.

See also BONE AND TEETH, PROPERTIES OF; COMPUTED TOMOGRAPHY.

BONE UNUNITED FRACTURE AND SPINAL FUSION, ELECTRICAL TREATMENT OF

AR LIBOFF
Oakland University
Rochester, Michigan

DEFINING THE UNUNITED FRACTURE

Within hours following a fracture in bone and the rapidly resulting hematoma, an endogenous repair process is initiated, characterized by increased cell division in the periosteum and endosteal stem-cell differentiation leading to organization of the hematoma into fibrocartilaginous callus. The latter represents the source of osteogenic potential from which ossification and subsequent bone remodeling occurs. In the ideal case, with proper management, those suffering bone fractures will normally find themselves fully recovered within a few months, with this time varying according to the specific bone involved, the type of fracture, and the age of the patient.

However, a small percentage of fractures fall outside the norm and do not heal as readily. There are upward of 5 million fractures occurring each year in the United States (1). Approximately 5–10% of these remain ununited after a few months. One can identify two types of ununited fractures, those undergoing *delayed* fracture healing, as evidenced by a lack of full healing in 3–6 months, and *nonunions*, where there is a lack of healing 6–12 months after the fracture has occurred. Marsh (2) suggests that the best measure of fracture healing in humans may be recovery of bending stiffness (i.e., the torque measured in Newton-meters that will bend bone by 1°). He defines delayed union as failure to reach a stiffness of of 7 $N \cdot m \cdot deg^{-1}$ at 20 weeks following fracture. Both the periosteum and the *end*osteum are deeply involved in the process of fracture healing, and it has been suggested (2) that delayed healing may be the result of cessation of the periosteal response before bridging has occurred, while nonunion may be indicative of a breakdown of both the periosteal and the endosteal repair mechanisms. A more general term for nonunion is pseudarthrosis, or false joint. Worth noting is that this problem is also infrequently found at birth (congenital pseudarthrosis). Electric and electromagnetic treatment is prescribed for both types of psuedarthroses, those that are the result of ununited fractures, and those that are found at birth. Further, because spinal fusion following back surgery can be problematic, electromagnetic treatment is also being used as an adjunctive procedure to promote spine fusion (3).

THE ELECTRIC CHARACTER OF BONE

Bone has a number of remarkable physical properties, particularly its electric character. Its electrical properties and the intimate relation of these properties to the growth process in bone were brought to light in a series of experimental discoveries, beginning in the 1950s. These revealed

1. A piezoelectric effect in bone.
2. A striking bioelectric signature specifically associated with developing bone.
3. A characteristic signature in adult unstressed bone.
4. A characteristic bioelectric signature following bone fracture.

Piezoelectricity is the rather unique property in which mechanical force is transformed into electric polarization (Fig. 1). Bone was shown to be piezoelectric by Yasuda in the early 1950s, but the work leading to this conclusion was not made generally available until 1957(4). Fukada and Yasuda (5) later found that this property could be traced to the intrinsic collagen component in bone. Since that time, a number of observers (6–8) suggested that this mechanical stress–electric polarization property should more properly be referred to as a stress-generated potential (SGP), reflecting the fact that what actually happens may not be the result of the special sort of crystal or textural structure that underlies the piezoelectric effect, but might instead result from the well-known electrokinetic effect of streaming potential. Streaming potentials, similar to piezoelectric signals, are characterized by the transformation of mechanical stress into a potential difference. However, streaming potentials do not occur because of any intrinsic crystal structure, but rather because fluid displacement through porous materials or tubes results in electric charge separation. It is generally agreed that dry bone indeed exhibits piezoelectricity, but opinions vary on whether this effect actually plays a role when bone is in its usual (i.e., wet) physiological environment. Part of the difficulty in resolving this issue is that the piezoelectric effect is not easily measured in wet bone. Whatever the pros and cons concerning studies on wet bone, it is difficult to put aside the seminal experiment by McElhaney (9). More than 600 silver epoxy electrodes were attached to cover the surface of a dried intact human femur from autopsy, and a vertical

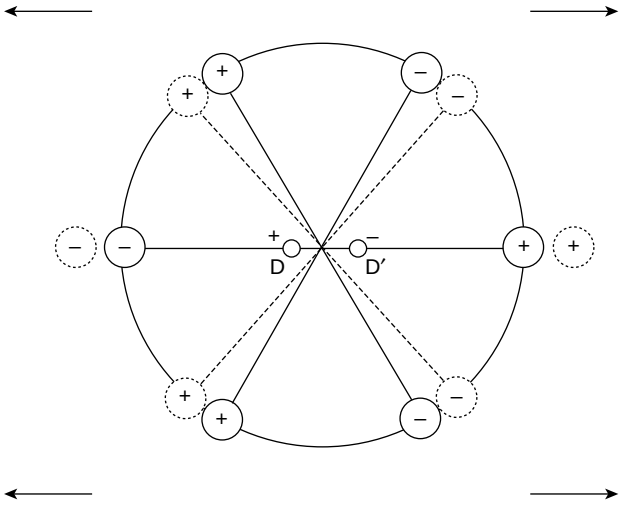


Figure 1. Piezoelectric Effect. Here, a tensile force results in a net electrical polarization in a material that ordinarily does not exhibit any polarization. Note that a compressive force will also result in electrical polarization. The source of the piezoelectric effect in bone is collagen.

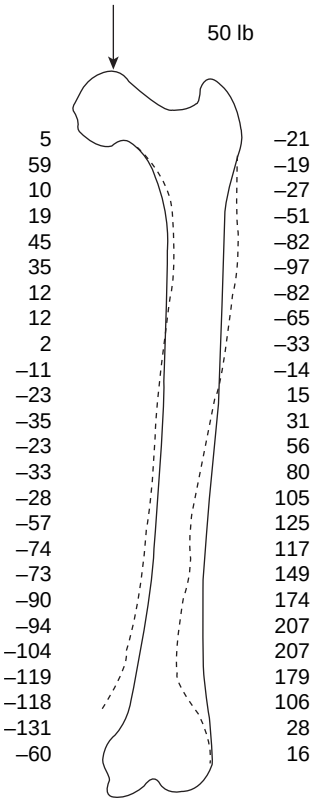


Figure 2. Map of piezoelectric voltages in dry stressed femur. When a 50-lb (220 N) load is impressed on dry human femur, piezoelectric voltages appear over the entire surface. One such set, given in millivolts (mV), is shown (9). The interpretation by Marino and Becker (10) is that the locus of these voltages, as shown by the dotted line for one slice through the femur, corresponds to the way that the bone will remodel under a specific load.

mechanical load was applied to the proximal end of the entire femur, mimicking the femur's weight bearing function. This load produced piezoelectric potentials from each of the electrode points, in effect mapping the piezoelectric response of the entire bone to this load. The voltages obtained varied widely in intensity, and included both positive and negative signs. These results were interpreted by Marino and Becker (10) as showing that, if one assumes that negative potentials tend to activate osteoblasts and positive voltages act to enhance osteoclastic function, then the voltage map (Fig. 2) represents the locus of the new remodeling surface for the femur: Areas of negative polarity are found where the femur needs thickening and areas of positive polarity are located where the bone must be reduced in thickness. Thus the potential remodeling response of the femur to the applied load is related in a very direct way to the polarity and intensity distribution of the piezoelectric signal. The McElhaney experiment showed convincingly that the piezoelectric effect in bone, *in the dry state*, conveys the information necessary to provide a remodeling template for bone under mechanical stress, in effect explaining Wolff's law (11), the empirical statement that bone remodeling follows the distribution of forces applied to the bone. Nevertheless, it is conceivable that the locus of voltages supplied by the piezoelectric effect in bone also requires local electrokinetic potentials to implement the remodeling process at the cellular level, either through cellular differentiation to produce the required osteoblasts and osteoclasts necessary for bone remodeling, or perhaps to separate the osteoblasts and osteoclasts by galvanotaxis (12).

Even in the absence of mechanical stress, bone exhibits a variety of intrinsic electric signals. One such effect is apparently part of the growth and development process. Measuring the electric potential in the same way for the same vertebral element from a group of cadavers covering a wide range of ages, Athenstaedt (13) found that this voltage was clearly connected to the age of the individual, greatest in infancy and ultimately falling to a level voltage plateau with maturity. The implication is that electric polarization in bone plays a role in the growth process. Something similar happens in long bone. One can measure voltage differences, usually referred to as bioelectric potential (BEP) along the length of a long bone (14) (Fig. 3). The BEP is always a relative measurement, where, for example, one can fix one electrode at one end (the epiphysis) and measure the potential difference at various points along the shaft of the bone (the diaphysis). Particular attention has focused on the growth plate, that region between epiphysis and diaphysis where the bone actually is ossifying as it grows. The BEP measured at the growth plate relative to the epiphysis in immature, growing, bone is markedly negative by as much as 5 mV (15), but as growth ceases, this potential difference becomes less pronounced. Furthermore, it has been demonstrated by means of tetracycline labeling (16) that the formation of new bone corresponds closely with the BEP profile.

Bone also exhibits an intrinsic electrical character even if it is not actively growing or under mechanical stress. For example, a BEP profile is also found in adult bone, albeit with a different signature. In measurements of this voltage, it is observed that the proximal metaphysis is always

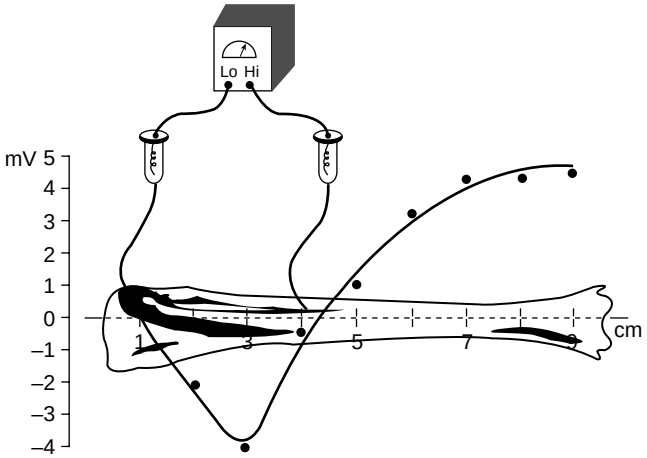


Figure 3. The BEP Profile. Bioelectric potential profile of a rabbit tibia (14). Voltage differences are obtained relative to the proximal end (on the left) by means of salt bridge electrodes.

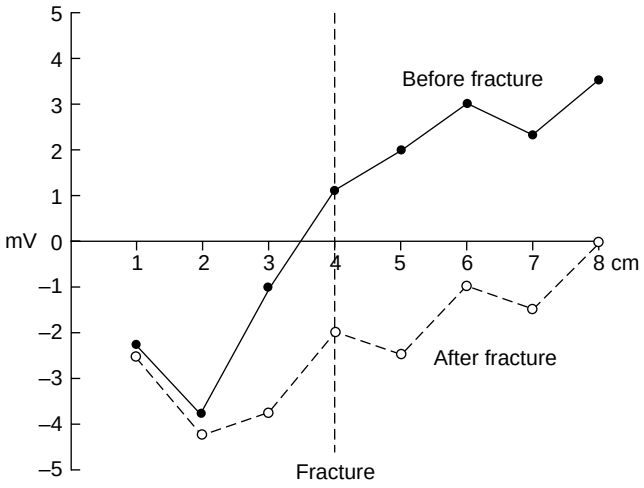


Figure 4. Effect of fracture of an 8 cm rabbit tibia on BEP. Measured potentials are shown before and after the fracture, which is at the 4 cm point.

negative with respect to the midshaft and distal portions of the bone. Because the BEP is unaffected by local nerve denervation or reduced blood flow, but slowly disappears following animal death, it is believed (17) that the origin of this voltage stems from functioning bone cells, acting in concert.

Other than this likely connection to bone cells, it is difficult to pin down a reasonable physical explanation for the ubiquitous potential profile associated with long bone. Electric polarization is readily observed in specimens of mature bone when they are even slightly heated. The origin of this effect is still unclear, but it may reflect a pyroelectric response having a textural origin (18), or perhaps, as Mascarenhas has suggested (19), bone is inherently an electret, a type of material, like many bioploymers, with the interesting property of being capable of storing electric charge. Electrets are the electrical equivalent of magnets, and some observers have suggested that bone exhibits ferroelectric properties. The characteristic property seen in electrets is a slow release of charge when heated. For example, long-term currents on the order of 100 fA can be observed (20) for bone specimens heated to 40 °C. Regardless of the cause, it is most likely the case, as stated by Brighton (21), that: *. . .in living, non-stressed bone, areas of active growth...[are] electronegative when compared with less active areas*.

There is one more impressive electric property associated with living bone, again reflecting this question of the role of negative potentials. Only a few hours following bone fracture, the bone becomes more negative relative to the prefracture BEP (22) (Fig. 4). There is some dispute as to whether this effect is limited to the fracture site or is distributed more widely along the length of the bone (14,23). This uncertainty is in all likelihood due to the fact that there are obvious measurement problems in obtaining a BEP profile for a fractured bone. As an injury current one might expect a more specific and localized expression. However, it is possible that the entire periosteum may be affected in a bone fracture at any point along its length. Worth noting are the experiments by Becker and Murray (24) on fracture healing in amphibian systems indicating a discrete electrical negativity at the fracture site, which led him to characterize the innate ability of bone to heal itself in higher animals as a form of regenerative healing.

Viewed in the context of its other electrical properties, the change in voltage profile associated with bone fracture has to be regarded as consistent with the overarching concept that bone makes extensive use of electricity in all of its growth, repair, loadbearing, and homeostatic processes. Because of this, it is hardly surprising that exogenous electric currents have been widely applied in attempts to grow and/or repair bone.

The FDA-approved devices for electric repair of ununited fractures fall either into invasive or noninvasive categories. The invasive devices make use of implanted direct current (dc) and (ac) electric signal sources, both pulsed and continuously sinusoidal. The noninvasive types are either purely electric (capacitive coupling or CC), or electromagnetic, using pulsed magnetic fields (PMF or PEMF) or ion cyclotron resonance (ICR) tuned magnetic field combinations.

DIRECT CURRENT OSTEOGENESIS

The surgeon who first observed that bone is piezoelectric, Iwao Yasuda, was also the first to demonstrate (25) that electric fields applied to long bone in vivo are capable of producing callus. He wrapped a few turns of wire around rabbit femur, and, maintaining this point at a negative potential, passed a small (1 μA) current to an anode located away from the bone. It was consistently observed that after 3 weeks this current resulted in spicules of osseus callus (called electric callus by Yasuda) (Fig. 5). Surprisingly, these spicules were not directed along the bone itself, but instead along the direction of the current, in some cases actually pointing away from the bone. To the orthopedic surgeon, one of the most positive signs during the course of fracture repair is the appearance of callus. Thus the observation by Yasuda cannot be overemphasized.

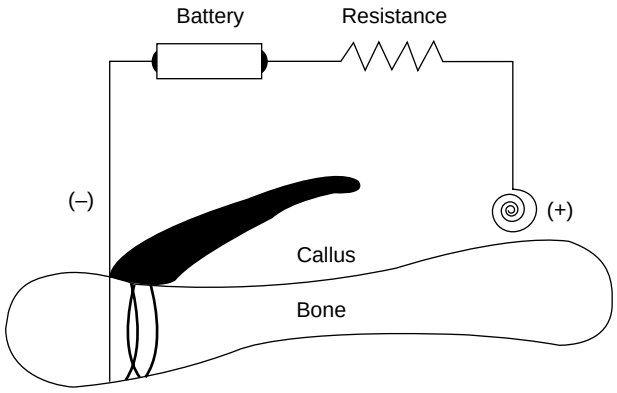


Figure 5. Formation of callus in response to 1 μA dc current. After original sketch by Yasuda (26), showing wires wrapped around the bone.

Although the reasons why electricity is capable of forming callus are still not clear, the implication of Yasuda's work was that electrical stimulation might be of assistance in bringing ununited fractures to closure.

Most of the follow-up experiments to Yasuda's discovery concentrated on determining the effects of electrical signals on normal and fractured bone. A commonly used animal experimental design was to apply an electrical signal to one femur while using the contralateral femur in the same animal as a control. Intrinsic to this approach was the use of dummy electrodes, carrying no current, but serving to affect the contralateral limb by its mere presence in whatever way the electrodes were affecting the activated side. It was in this manner that Bassett et al. (27) used implanted battery packs to deliver microampere-level currents to platinum–iridium wire electrodes extending into the medullary cavities of femora in dogs. The results clearly indicated that more bone was formed in the intermedullary space in the vicinity of the cathodes than near the anodes. Follow-up experiments (28) reinforced the finding that bone growth appeared to be effective at the cathode, but also that bone necrosis occurred at anodes for currents in excess of 20 μA. In this work Friedenberg et al. (28) found that bone growth was most pronounced for currents between 5 and 20 μA. There is some question concerning this optimal current in that the required levels may be dependent on the mode of application. In rabbit femur, circular defects ~ 2.8 mm in diameter were repaired within 3 weeks when subjected to currents ranging between 2.5 and 3 μA applied by two electrodes on either side of the defect (29). Not only was the current lower than that suggested by Friedenberg et al. (28), but there was no particular advantage to either polarity. Similarly, Hambury et al. (30) studying $_{85}$Sr uptake in rabbit femur observed osteogenesis at 3 μA, again with no difference due to polarity. In another attempt (31) to establish the optimal current for repairing bone defects in dog, it was reported that 0.2 μA was more effective than either 2.0 or 20 μA.

Further complicating the issue of what level of current is required to initiate osteogenesis were a number of earlier reports in which callus was formed using currents that were orders of magnitude smaller than microampere levels. Fukada and Yasuda (4) wrapped a charged Teflon electret around bone to initiate callus, work that was later successfully repeated in Japan (32,33). Three different types of current application were employed in the latter experiment: that emitted by an electret, that obtained from the piezoelectric poly-γ-methyl-L-glutamate (PMLG) film, and a battery delivering 8–10 μA. The two current levels for the electret and the film, respectively, were 1 and 10 pA, levels smaller by huge factors of 10^{-7} and 10^{-6} from the "optimal" value of 10 μA. Marino and Becker (34) raised the issue as to whether this enormous difference in currents, both seemingly effective, means that more than one mechanism is involved, with the microampere (μA) results indicative of a nonspecific osteogenic stimulus while the picoamp (pA) currents more closely mimicing the endogenous piezoelectric response.

ELECTROMAGNETIC OSTEOGENESIS

Among his other important discoveries, Michael Faraday was the first to show that voltage is induced in a conductor when a nearby magnetic field is changing rapidly. This phenomenon, often referred to as Faraday's law, can be mathematically expressed by the following expression:

$$dB/dt = -V/A \tag{1}$$

where dB/dt is the time rate of change of the magnetic field B through a region of area A, and V is the voltage induced by dB/dt along the path that is circumferential to A by this rate of change. If B is varying at some frequency f, the product fB is a good measure of the relative effectiveness of dB/dt. When the region in question is electrically conducting, as in living tissue, one can use Ohm's law to rewrite the above expression in terms of the current I instead of V. Thus if R is the resistance of the circumferential path around A, Eq. 1 is changed to read

$$dB/dt = -I(R/A) \tag{2}$$

In this way, one can induce a current in the vicinity of a bone defect by employing a nearby magnetic field that is changing rapidly—the faster the rate of change, the greater the current. One achieves a faster change (i.e., a larger dB/dt) by merely increasing the frequency at which B is changing. Further, it is important to realize that the current so induced is no different from currents that are produced by purely electrical means (Fig. 6). Most important, since the source of the magnetic field can be deployed externally, Faraday's law enables the clinician to generate the required therapeutic currents in a completely noninvasive manner.

In 1974, following 5 years of intensive effort. Bassett and Pilla(35) reported on the successful use of the Faraday induction concept (PMF) to repair fibular osteotomies in beagle. Coils were placed on either side of the leg in such a manner that the magnetic fields from the coils traversed the defect and were additive (Fig. 7). The currents through each coil were pulsed in two ways, at 1 pulse $\cdot$ s^{-1} and at 65 pulses $\cdot$ s^{-1}. As revealed by mechanical testing of the fibula subsequent to treatment, there was greater indication of recovery with 65 pulses $\cdot$ s^{-1}, a finding that was consistent

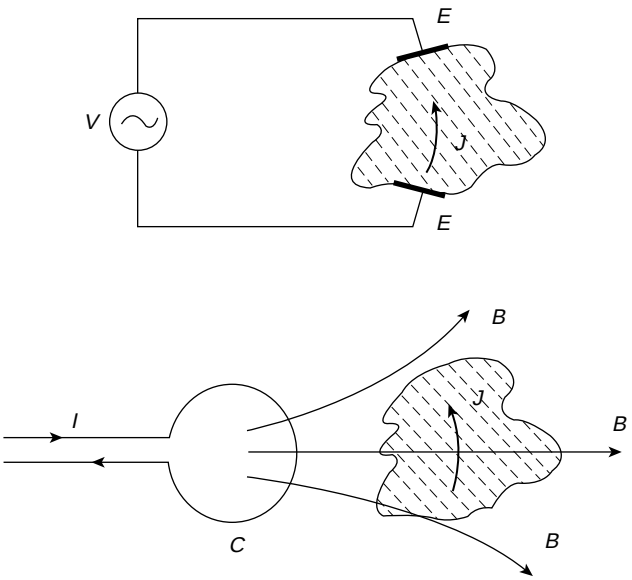


Figure 6. The current density J produced in tissue by a voltage source V acting through electrodes E is no different from the current density induced by a changing magnetic field B according to Faraday's law. The B field is produced by a current I that energizes **a** coil C, whose plane is perpendicular to the page.

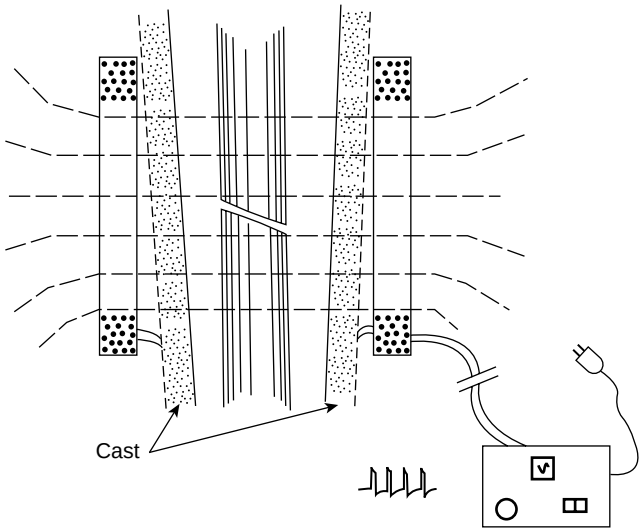


Figure 7. The PMF technique uses two flat coils connected in series to generate a magnetic field (dashed line) through the bone defect. Because the magnetic field is changing rapidly, a voltage is induced, producing a current in the vicinity of the defect. The process is completely noninvasive. In this sketch (36) the two parallel coils, whose planes are perpendicular to the page, are shown outside the cast.

with the prediction from Eq. 2 of a larger current with higher frequency.

INVASIVE (IMPLANTED) ELECTRIC TREATMENTS

Although the use of pulsed magnetic fields provides a means by which one avoids electrode implantation, some surgeons still prefer the extra advantages that come with direct observation of the pseudarthrosis defect. In addition, there is a very lengthy literature background on delivering dc directly to bony defects.

In late 1971, groups at New York University (NYU) and at the University of Pennsylvania independently demonstrated that electrical stimulation using implanted dc devices was successful in repairing pseudoarthrosis defects in humans. In both cases, electrodes and battery were surgically implanted with provisions for percutaneously monitoring the current. Otherwise, however, the methods employed were strikingly different. The NYU group, led by L.S. Lavine (37) used platinum wire electrodes on either side of a congenital pseudarthrosis in the lower tibia of a 14 year old male, in effect allowing the current to pass through the defect (Fig. 8). The polarity of the current was such that the proximal side of the defect was negative. This approach was the same as successfully used in this group's previous experiment (29) to repair defects in rabbit femur. The current was monitored and maintained over the 18-week treatment period at 3.9 μA.

By contrast, Friedenberg et al. (38) in treating a nonunion in the medial malleolus of a 51 year-old woman, used a technique (Fig. 8) that had been previously been found to be successful in producing callus in rabbit fibula (28,39). A stainless steel cathode was located directly in the defect and the anode, an aluminum grid, was taped to the skin. A constant-current power source maintained the current at 10 ± 2 μA over the 9 week treatment period. Again, as with the nonunion treatment employed by the NYU group, the outcome was successful.

These differences in treatment, both leading to repair of the nonunions, remain unresolved. The one treatment (37) is consistent with prior animal work in which the proximal side of bone was found to be intrinsically negative, while the second result (38) fits those observations (24) claiming that fractures are more negative than the rest of the bone. These differences tend to highlight a key difficulty connected to the research on the electric treatment of bone. Apart from the essentially empirical nature of measurements such as the BEP profile, there is no fundamentally sound basis with which to explain the underlying mechanisms, resulting in continuing uncertainties in the clinical techniques.

A number of investigators have attempted to shed light on this question of mechanisms. Almost all such "explanations" have focused on the electrically related regulation of different factors: parathyroid hormone (PTH) (40), adenosine 3', 5'- monophosphate (cAMP) (41,42), insulin-like growth factor II (IGF-II) (43), bone morphogenetic protein (BMP) (44), transforming growth factoř-beta 1 (TGF-β1) (45), and calcium ion channel transport (46). These are contributors, in varying degrees, to the cellular signal transduction pathways controlling bone growth. However, it would be truly surprising if these factors were *not* involved in all types of osteogenic processes, including electrical osteogenesis. At best, such factors must be regarded as merely *indicators* of metabolic activity in bone. At this point in time, they provide little, if any clue as to the reason why bone is responsive to electrical stimulation.

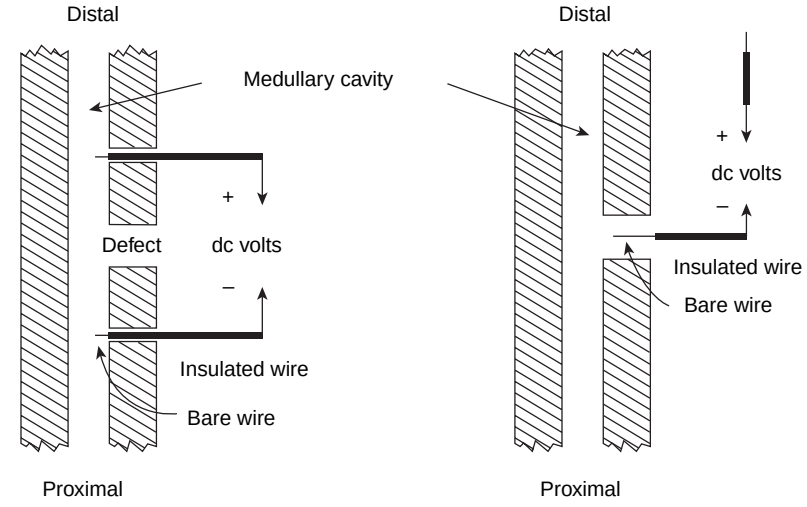


Figure 8. Two ways of using μA-level dc currents to repair defects in bone. In one case, the electrodes are applied so as to bridge the defect. In the other case, the cathode is placed directly into the defect. Both approaches have been successful (37,38) in treating human nonunions.

Presently, there are two FDA-approved implantable direct current devices for treating bony defects, both marketed by ElectroBiology Inc. (Parsippany, NJ). These are shown in Figs. 9 and 10. The Osteogen Bone Growth Stimulator supplies 40 μA through mesh electrodes. Although the cathode is located at the defect, similar to the original placement by Friedenberg et al. (38), the current is far in excess of what was thought to lead to bone necrosis (28). Apparently, the nature of the electrodes used by Friedenberg et al. (28) may have played a role in this discrepancy. The second implantable dc device is the SpF Spinal Fusion Stimulator, prescribed for spinal fusion. The positive and negative leads, in this case carrying 60 μA, are located on either side of the repair site.

NONINVASIVE ELECTRIC TREATMENT: (CC)

One method for applying an electric current to a defect in bone in a noninvasive manner is by means of capacitive coupling (CC). The background for this technique were experiments (47,48) in which 60 kHz sinusoidal voltages were capacitively transferred to bone cell cultures resulting in an electric field within the culture medium of 20 $mV \cdot cm^{-1}$ and a current density of 300 $\mu A \cdot cm^{-2}$. The first clinical use of this was to treat nonunions (49) (Fig. 11). An overall efficacy of 77% was achieved in a group of 22 cases with a mean time to healing of 23 weeks. The FDA-approved version of this technique is marketed by EBI (Fig. 12) As in the earlier studies on cell culture, the pictured device makes use of a 60 kHz alternating electric field that is applied to the skin on either side of the defect using disk electrodes and conducting gel. The current density within the tissue is considerably less than the levels used in cell culture, only $\sim 7\ \mu A \cdot cm^{-2}$. The term *capacitive* may not be warranted for the devices pictured in Figs. 11 and 12. Unlike the lack of ohmic coupling in the earlier, *in vitro* studies, there is a much larger ohmic contribution to the overall impedance when disk electrodes are used. A better description for this device category might be ac (i.e., simply alternating current) instead of CC.

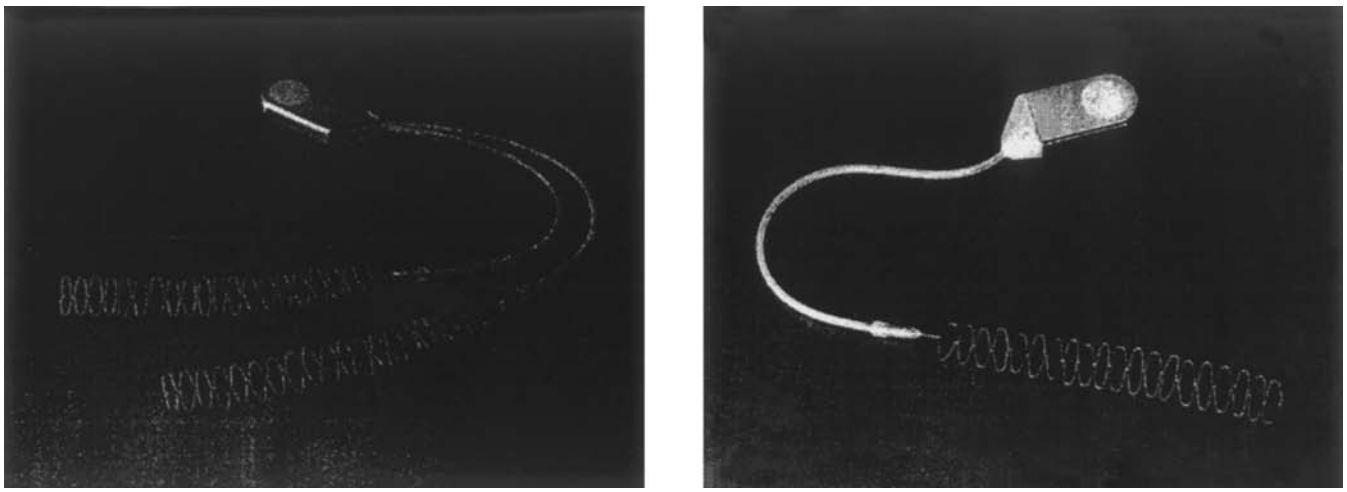


Figure 9. Implantable device for bone growth stimulation. EBI (Electro-Biology, Inc.)

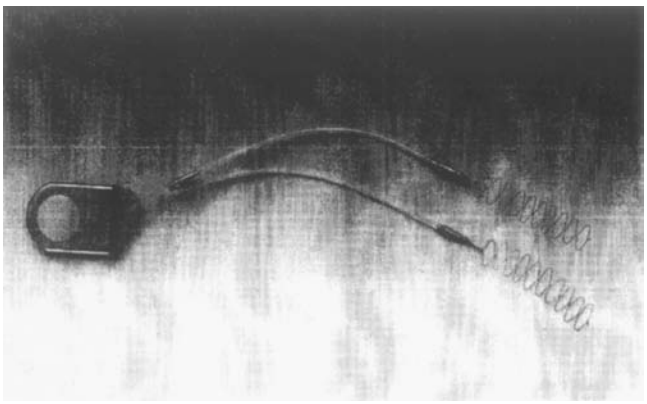


Figure 10. Implantable device for adjunctive treatment of spinal repair (EBI)

A wide range of parameters have been used in studying the clinical and experimental aspects of the CC signal, with various voltages applied to the skin between 1 and 10 V, and frequencies between 20 and 200 kHz. The electric field strengths generated within tissue has ranged from 1 to 100 $mV \cdot cm^{-1}$ and the current densities from 0.5 to 50 $\mu A \cdot cm^{-2}$.

NONINVASIVE ELECTROMAGNETIC DEVICES: (PMF)

The successful use of PMF (also called PEMF) by Bassett et al. (35) to repair bone defects in animals noninvasively led to a number of different devices aimed at applying pulsed magnetic fields. One such early design, successfully applied to the treatment of a tibial nonunion (50,51), made use of an iron-cored electromagnet driven by a square pulse with a repetition rate of 1 pps. However, this design suffered because the large inductive reactance of the iron core acted as a constraint on the repetition rate of the coil current pulses.

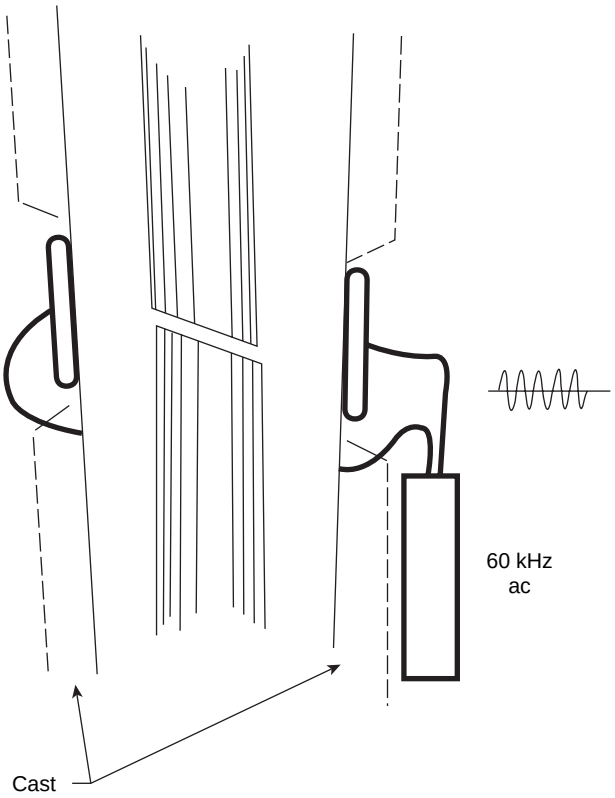


Figure 11. Capacitive Coupling. Electrodes are attached on either side of the bony defect external to skin (here, external to cast) supplying a 60 kHz sinusoidal signal.

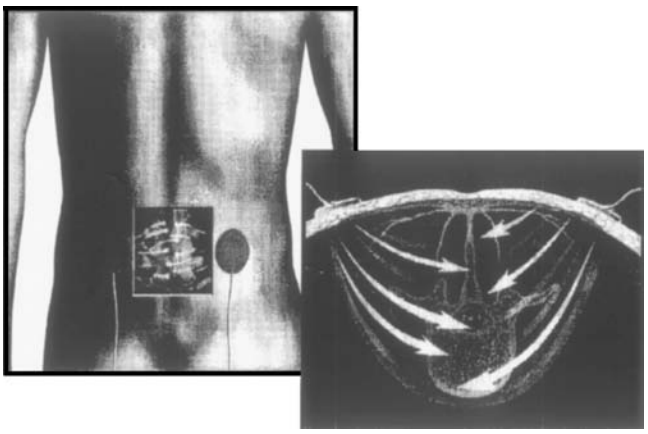


Figure 12. Mode of action of EBI capacitive coupling spinal fusion stimulator (EBI).

With this constraint in mind, Bassett's group succeeded in designing (52) a low inductance air-coil system that could be pulsed at higher frequencies to repair recalcitrant pseudarthroses and nonunions in humans (Figs. 13–16). The success rate that was reported (85%) was greatly in excess of the salvage rate usually obtained by orthopedists using conventional, nonelectrical procedures. However, later (55), reviewing PMF treatments for a wider, all-inclusive group of pseudarthrosis cases, including those with the worst prognosis, Bassett lowered the success rate downward, to 54%.

The pulsed magnetic field that was originally used by Bassett was (and still is) based on the saw-tooth signal common to the fly-back refresher circuit in television receivers. A saw-tooth voltage (Fig. 17) is applied to a pair of many turn coils, creating a current in both coils that generates a single magnetic field. The planes of the coils are roughly parallel, and deployed on opposite sides of the defect (see Fig. 7), creating a commonly directed magnetic field through the defect. The sawtooth signal applied to the coils results in a rapidly changing magnetic field, ~ 10 tesla per second ($T \cdot s^{-1}$), maximized at those times when the voltage applied to the coils is falling sharply. Faraday's law results in the induced voltage shown in Fig. 18. The net induced signal that appears in the vicinity of the defect consists of bursts of 21 pulses, each individual pulse 260 μs in duration, with the bursts repeating at 15 Hz. The magnetic field that actually appears in the area of the defect rises, with each pulse, to ~ 10 G (1 mT), before dropping precipitously in ~ 25 μs. It is this rapid change in B that contributes the most to the induction of current (see Eq. 2). For example, if a sine wave signal of 10 G at 60 Hz were applied to the same region instead of this pulse, the maximum current would be 600 times smaller.

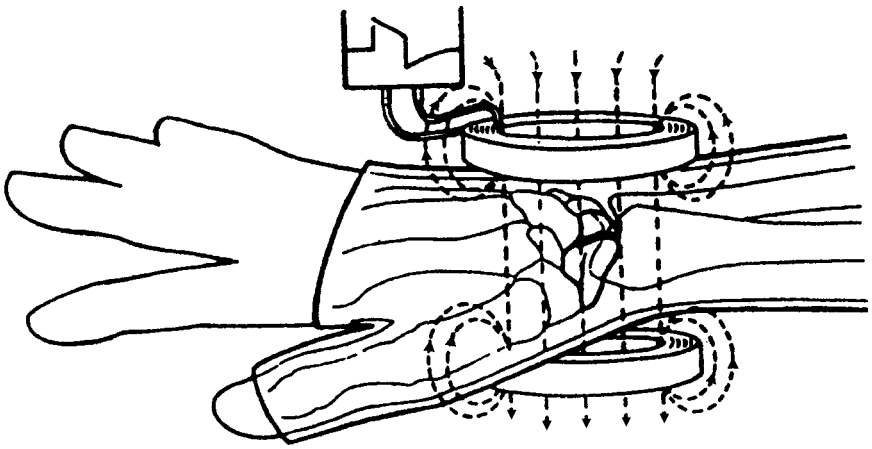


Figure 13. Treatment of ununited scaphoid fracture with PMF (53).

The various PMF clinical and experimental signal repetition rates that have been attempted vary between 1 and 100 Hz, with the maximum magnetic field intensity at the defect site ranging from 0.1 to 30 G, and the induced electric field at the site ranging between 0.01 and 10 $mV \cdot cm^{-1}$.

NONINVASIVE ELECTROMAGNETIC DEVICES (ICR)

Magnetic fields are also used in bone repair in ways that have nothing to do with Faraday induction. It was shown in 1985 (56) that the results embodied in the so-called calcium efflux effect (57,58) were in close agreement with predictions based on the resonance characteristics of certain biological ions subject to the Lorentz force. Specifically, the shape of the nonlinear frequency dependence of calcium binding to chick brain tissue was what might be expected for a particle with the charge-to-mass ratio of the potassium ion moving in combined parallel sinusoidal and dc magnetic fields whose ac frequency and dc intensity corresponded to the ICR condition for K^+. This observation also explained earlier work (59,60) in cell culture demonstrating that weak low frequency magnetic fields enhance DNA synthesis in a manner that is clearly not related to Faraday induction, since the additional DNA synthesis does not scale with either frequency or intensity. Ion cyclotron resonance is a magnetic effect that is fundamentally different from Faraday's law as expressed in Eq. 1. More specifically, as regards possible effects of magnetic fields on bone, it entails a totally different phenomenon than the induction of current in bone using pulsed magnetic fields.

Unlike previous attempts to arrive at the electromagnetic conditions required for electrical osteogenesis, exact predictions are possible using the ICR effect. One can focus on a specific ion and adjust the intensity of the dc magnetic

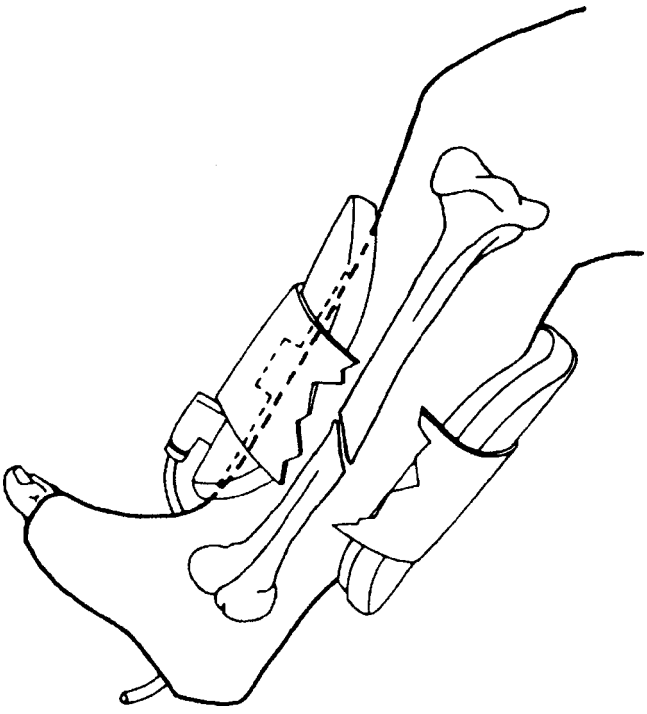


Figure 14. Treatment of congenital pseudarthrosis of the tibia with PMF (54).

Figure 15. PMF Bone healing system (EBI).

Figure 16. PMF device in place on patient (EBI).

field and the frequency of the ac magnetic field to "tune" to this ion. This is because a resonant condition occurs when the ratio of the frequency of the ac field to the intensity of the dc field is equal to the charge-to-mass ratio of the ion. The simple expression governing this resonance is

$$\omega/B = q/m \tag{3}$$

where ω is the (angular) frequency of the ac field, in $\text{rad} \cdot \text{s}^{-1}$, B is the intensity of the dc magnetic field, in Tesla, and, q/m is the mass-to-charge ratio of the ion. For practical applications, the angular frequency ω is replaced by its equivalent, $2\pi f$, where f is the frequency in hertz (Hz). The underlying interaction mechanism for this effect in living tissue is still in question (61), but the most reasonable explanation is that ions in resonance are more likely to stimulate the gating mechanism for ion channel transport.

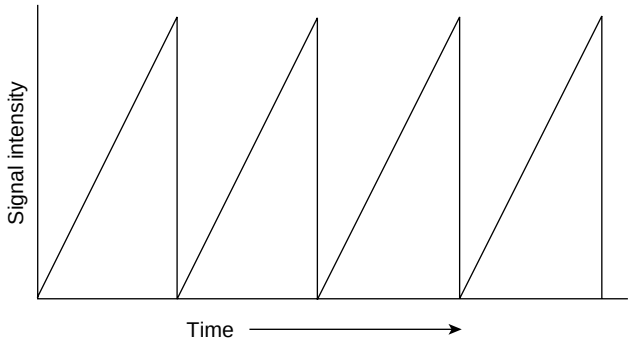


Figure 17. Sawtooth voltage applied to PMF coil.

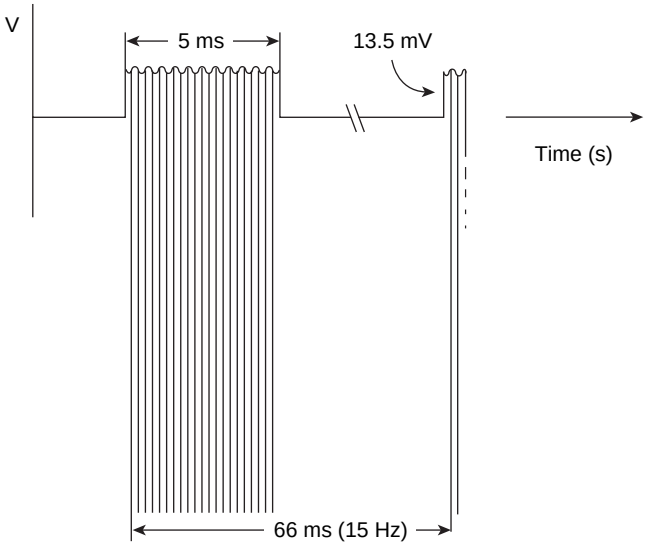


Figure 18. Voltage induced in tissue by PMF.

A great deal of work has been done in examining the effects on biological expression when tuning to Ca^{2+}, Mg^{2+}, and K^{+}, not only in bone cell culture (Fig. 19) (62), but also in neural cell culture, in animal behavior, and in plants (61). It is generally agreed that ICR tuning to these ions can have striking effects on growth. One such example (63) is shown in Fig. 20 illustrating the relative effects on explanted embryonic chick femora cultured under Ca^{2+} and under K^{+} ICR magnetic field conditions.

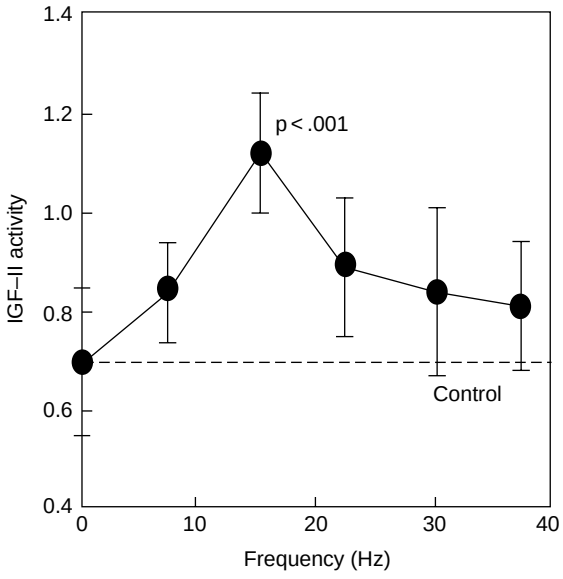


Figure 19. Frequency response of insulin-like growth factor in bone cell culture under combined ac and dc magnetic field exposure 62. The dc field was maintained at 20 μT for each of the points shown. There is a clear peak at 15.3 Hz, corresponding to the predicted ICR condition for Ca^{2+} resonance in Table 1.

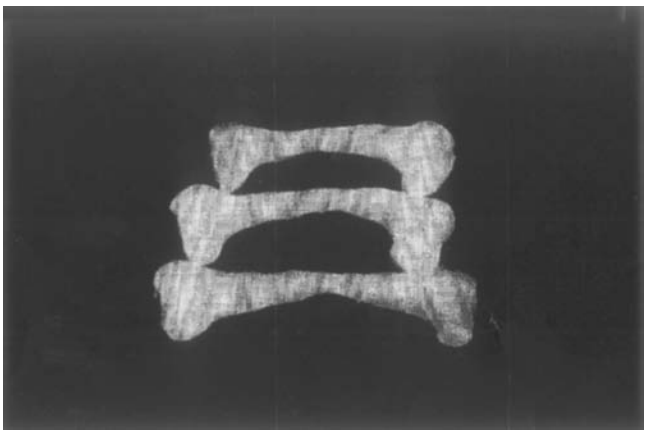


Figure 20. Effect of ICR magnetic exposures on chick embryonic growth (63). The topmost femur, the shortest, was grown under K^+ ICR tuning, while the bottom femur was grown under Ca^{2+} ICR magnetic field conditions. The middle femur was not exposed to any ICR field.

Diebert et al. (64) examined the efficacy of ICR in repairing defects in rabbit fibula, basically reusing the animal model that had been previously employed to study electrical osteogenesis (29), but applying an ICR magnetic field combination instead of a dc current. It was found that the 28-day ICR treatment yielded results equivalent to or better than those employing direct current and pulsed magnetic fields. For animals exposed to Ca^{2+} resonance magnetic fields for as little as 30 min · day^{-1}, there was an average increase in stiffness of 175% over controls, rising to nearly 300% when the exposures were maintained for 24 h. Somewhat smaller increases in stiffness were also observed for exposures tuned to the Mg^{2+} charge-to-mass ratio.

Another aspect of the ICR effect is that one can also use harmonics, that is, multiples of the frequency condition given in Eq. 3. For theoretical reasons (65) only odd harmonics are allowed. Thus, the most general expression for cyclotron resonance frequencies is

$$f_n = (2n + 1)(1/2\pi)(qB/m) \quad n = 0, 1, 2, 3, \ldots \qquad (4)$$

Table 1 lists the frequency/field ratios (f_n/B) for the three ions, Mg^{2+}, Ca^{2+}, and K^+ for the first three harmonics from Eq. 4. Note that some of these ratios are numerically close to one another. The 5th harmonic of Ca^{2+} is slightly $> 1\%$ greater than the 3rd harmonic for Mg^{2+} (3.83 vs. 3.79). This observation led S.D. Smith to suggest that using a frequency/field ratio of 3.8 might be particularly effective in bone where growth is indicated for both Ca^{2+} and Mg^{2+} stimulation (66). This ratio is the basis for a number of bone stimulation devices manufactured by the djOrthopedics

Table 1.

Ion	Fundamental f_0/B, Hz · μT	3rd Harmonic f_1/B, Hz · μT	5th Harmonic f_2/B, Hz · μT
Mg^{2+}	1.26	3.79	6.31
Ca^{2+}	0.77	2.30	3.83
K^+	0.39	1.18	1.97

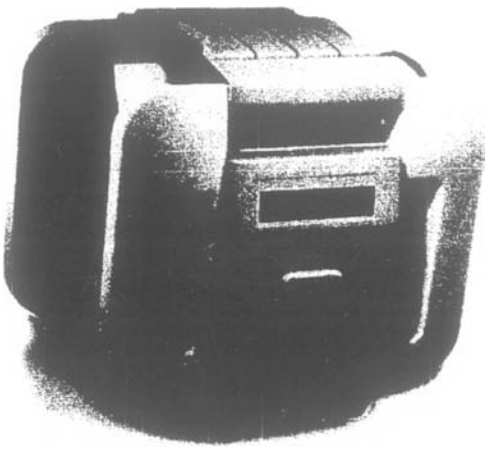


Figure 21. ICR bone repair device for treating nonunions. (djOrthopedics Corp.)

Corporation for treating pseudarthroses and enhancing spinal fusion (Figs. 21,22). The time variation of the magnetic field generated by these devices is shown in Fig. 23. Because the ac and dc magnetic field directions must be maintained parallel to ensure the resonance condition, these clinical devices achieve the frequency/field ratio by fixing the frequency of the applied sinusoidal magnetic field at 76.9 Hz, while using a second coil to continuously adjust for changes in the parallel component of the local dc magnetic field, to maintain this dc level at 20 μT.

Some observers incorrectly use the term *combined magnetic field* (CMF), to characterize this clinical technique. It is important to understand that the fields that are combined are highly specific, following the rules expressed in Eq. 4. In addition, it is possible to achieve the same conditions in tissue with a single magnetic field, using a prepared current derived from an arbitrary waveform generator. For these reasons, the term ICR should be used for all clinical and research techniques that are otherwise termed CMF.

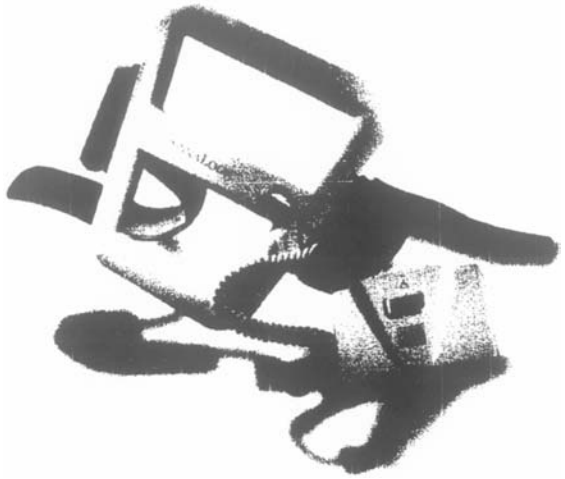


Figure 22. ICR device for adjunctive use in spinal fusion. (djOrthopedics Corp.)

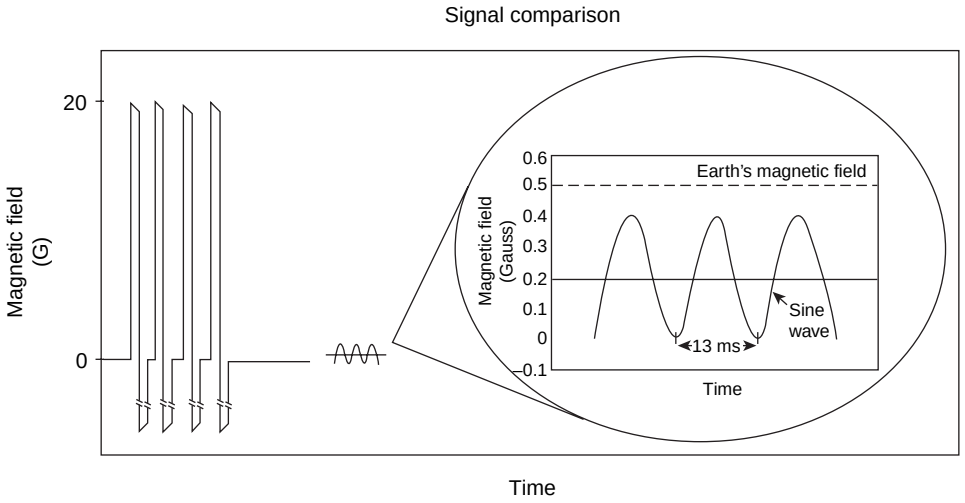


Figure 23. Comparing ICR signal to PMF signal. In the one case, there is a 20 μT peak sinusoidal magnetic field, and in the other a very short magnetic pulse ~ 100 times larger in intensity (Orthologic Corporation).

It is also sometimes incorrectly reported that ICR is an *inductive* procedure. However, the inductive current generated in the ICR device is negligible, approximately a factor of 10^{-5} smaller than the currents induced by PMF devices. While clearly noninductive, the actual ICR interaction mechanism is still in question (45). It is most likely coupled to events occurring at membrane bound ion channels (40), as evidenced that the calcium channel blocker nifedipene prevents the ICR response (67). It has been suggested, in this regard, that the channel gating process may be sensitive to the resonance tuning of specific ions (45).

SUMMARIZING EFFICACIES FOR THE VARIOUS TREATMENT

The three types of noninvasive treatments for bone defects, PMF, ICR, and CC, have each been subjected to randomized, double-blind trials and are shown to be efficacious, with an overall success rate of between 50 and 70%. One reason for this variation is undoubtedly the inclusion, or lack therein, of patients with defects that are intrinsically more difficult to repair. As the gap in a pseudarthrosis extends to widths > 5 mm, the likelihood of successful treatment diminishes. For this reason, some clinicians choose to exclude patients with radiographic gaps > 5 mm from electrical treatment (21,68).

More than 20 years after Bassett's original use (52) of pulsed magnetic fields to repair nonunions, a definitive work on using PMF to treat delayed unions was published by Sharrard (69). A total of 45 fractures of the tibia were examined in a double-blind multicenter trial, with active PMF stimulation in 20 patients and dummy control units in 25 patients for 12 weeks at 12 h/day. The results, 9 unions in the active group compared to only 3 in the control group, were "*very significantly in favour of the active group* ($p = 0.002$)". This effectiveness of PMF stimulation was confirmed for the case of tibial osteomoties in still another randomized, double blind study (70). Similarly, the successful use of CC and ICR, respectively, in treating nonunions, was reported by Scott and King (71) and by Longo (72). Recently, there has been increasing interest in the use of these electromagnetic techniques as an adjunctive to spine fusion. Again, as with the treatment of pseudarthroses, randomized, double-blind trials carried out for the PMF (73), CC (74) and ICR (75) techniques, have indicated that each is also efficacious in the adjunctive treatment of spine fusion.

GENERAL REMARKS

A summary of the electrical and electromagnetic treatments for nonunions and spinal fusion is given in Table 2. It is difficult to make simple comparisons based solely on the relative electrical characteristics, since each modality is based on different types of specifications, including current, current density, time rate of change of magnetic field, frequency, and magnetic intensity. As mentioned above, the levels of current that have been used to achieve osteogenesis extends over a range that has many orders of magnitude, a fact that seems to preclude any mechanism that is simply connected to current alone. It is highly likely that the larger of these successful current levels achieve osteogenesis, as Becker has suggested, by acting as an irritant, and that the smaller levels are perhaps related to the sorts of currents that might occur naturally, perhaps as the result of stress-generated potentials. One measure of

Table 2. Summary of Electrical and Electromagnetic Treatments for Nonunions and Spinal Fusion

	Modality	Designation	Characteristics	Daily Treatment
Invasive	dc electric	dc	1 pA–60 μA	24 h
Noninvasive	ac electric	CC	60 kHz, 7μA · cm^{-2}	24 h
	Pulsed magnetic field	PMF, PEMF	dB/dt=100 T · s^{-1} Repetition rate=15 Hz	3 h
	Sinusoidal magnetic field	ICR, CMF	77 Hz ac frequency 20 μT dc Field	30 min

this potential dichotomy is the remarkable fact that both ICR and PMF techniques are equally successful in treating nonunions, despite the fact that the induced currents differ by a factor of 10^5.

There is undoubtedly room for improvement in the efficacy of the various electromagnetic treatments to repair bony nonunion. Note that both treatments, PMF and ICR, were each adopted for clinical use on the basis of the original designs, with no subsequent studies before or after FDA approval that might have been initiated to search for waveforms and signals that conceivably could be used to optimize treatment. Thus for pulsed magnetic fields, it remains to be seen what roles are played by variables such as pulse width, rise time, repetition rate, and so on, and whether marked improvements in efficacy would follow optimization of these key variables. At least one report (76) claims that peak magnetic fields 100 times smaller than used in the EBI PMF device are just as effective in treating nonunions. Similarly, positive results were obtained in treating tibial osteotomies in rabbit with very different pulse characteristics from that of the EBI clinical device (77). Not only was the magnetic pulse reduced by a factor of 15, but the pulse repetition rate was reduced by a factor of 10, and the frequency components in excess of 20 kHz were filtered from the signal. This lack of optimization is equally true for the djOrthopedics ICR therapeutic signal, based on an approximate simultaneous stimulation of Ca^{2+} and Mg^{2+}ions as well as a very specific ratio of ac to dc magnetic intensities. The ICR device presently approved by the FDA sets this ratio at unity, despite the fact that a number of investigators (78–80) suggested that this ratio may have important consequences for the efficacy of the resonance interaction.

Furthermore, it has been suggested (31,81) that the fundamental reason why some electrical treatments of pseudarthroses are successful may have little to do with the nature of the electrical signal itself, but rather that the initiation of callus formation is known to be tied to local irritants, such as occurs with mechanical, thermal, or chemical sources. It is not inconceivable that the efficacy of treatments such as PMF may result from its role as an irritant. There is evidence (82,83) indicating an increased expression of heat shock proteins in response to low level electromagnetic fields. This type of genetic expression can result from a wide range of stress factors.

The fact that electrical osteogenesis occurs naturally, in growth, homeostasis, and repair and, further, that it can be brought about by exogenous application, begs the question as to whether the present 50–70% repair rate might be substantially improved with further research into the actual underlying mechanism.

BIBLIOGRAPHY

Cited References

1. Ryaby JT. Clinical effects of electromagnetic and electric fields on fracture healing. Clin Orth Rel Res 1998;355S: 205–215.
2. Marsh D. Concepts of fracture union, delayed union, and nonunion. Clin Orthop 1998;355S:22–30.
3. Marone MA, Feuer H. The use of electrical stimulation to enhance spinal fusion. Neurosurg Focus 2002;13: article 6.
4. Fukada E, Yasuda I. On the piezoelectric effect in bone. J Phys Soc Jpn 1957;12:1158–1162.
5. Fukada E, Yasuda I. Piezoelectric effects in collagen. Jpn J Appl Phys 1964;3:117–121.
6. Anderson JC, Eriksson C. Electric properties of wet collagen. Nature (London) 1968;218:166–168.
7. Anderson JC, Eriksson C. Piezoelectric properties of dry and wet bone. Nature (London) 1970;227:491–492.
8. Pienkowski D, Pollack SR. The origin of stress generated potentials in fluid-filled bone. J Orthop Res 1983;1:30–41.
9. McElhaney JH. The charge distribution on the human femur due to load. J Bone Joint Surg 1967;49:1561–1571.
10. Marino AA, Becker RO. Piezoelectric effect and growth control in bone. Nature (London) 1970;228:473–474.
11. Wolff J Das Gesetz der. Transformation der Knochen. Berlin: A. Hirschwohl; 1892.
12. Ferrier J, Moss SM, Kanehisa J, Aubin JE. Osteoclasts and osteoclasts migrate in opposite directions in response to a constant electrical field. J Cell Physiol 1986;129:283–288.
13. Athenstaedt H. Permanent electric polarization and pyroelectric behavior of the vertebrate skeleton III. The axial skeleton of man. Z Zellforsch 1969;93:484–504.
14. McGinnis ME. The nature and effects of electricity in bone, Chapt. 6 in Electric Fields in Vertebrate Repair. In: Borgens RP, Robinson KR, Venable JW, Jr, McGinnis ME, editors. New York: Alan R. Liss; 1989.
15. Friedenberg ZB. Bioelectric potentials in bone. J Bone Joint Surg 1966;48A:915–923.
16. Rubinacci A, Tessari L. A correlation analysis between bone formation rate and bioelectric potentials in rabbit tibia. Calc Tiss Res Int 1983;35:728–731.
17. Friedenberg ZB, Harlow MC, Heppenstall RB, Brighton CT. The cellular origin of bioelectric potentials in bone. Calc Tiss Res 1973;13:53–62.
18. Lang SB. Thermal expansion coefficients and the primary and secondary pyroelectric coefficients of animal bone. Nature (London) 1969;224:798–799.
19. Mascarenhas S. The electret effect in bone and biopolymers and the bound-water problem. Ann NY Acad Sci 1974;238: 36–52.
20. Liboff AR, Furst M. Pyroelectric effect in collagen and structures. Ann NY Acad Sci 1974;238:26–35.
21. Brighton CT. The treatment of non-unions with electricity. J Bone Joint Surg 1981;63A:847–851.
22. Friedenberg ZB, Smith HG. Electric potentials in intact and fractured tibia. Clin Orthop 1969;63:222–225.
23. Becker RO, Spadaro JA, Marino AA. Clinical experiences with low intensity direct current stimulation of bone growth. Clin Orthop Rel Res 1977;124:75–83.
24. Becker RO, Murray DG. The electric control system regulating fracture healing in amphibians. Clin Orthop Rel Res 1970;73:169–198.
25. Yasuda I. Fundamental aspects of fracture treatment. J Kyoto Med Soc 1953;4:395–406 (in Japanese) Translated in *Clin Orthop* 1977;124:5–8.
26. Yasuda I. Mechanical electrical callus. Electrically Mediated Growth mechanisms in Living Systems. Ann NY Acad Sci 1974;238:457–465.
27. Bassett CAL, Pawluk RJ, Becker RO. Effect of electric currents on bone in vivo. Nature (London) 1964;204:652–654.
28. Friedenberg ZB, Andrews ET, Smolenski BI, Pearl BW, Brighton CT. Bone reaction to varying amounts of direct current. Surg Gyn Obstet 1970;127:894–899.
29. Lavine L, Lustrin I, Shamos MH, Moss ML. The influence of electric current on bone regeneration in vivo. Acta Orthop Scand 1971;42:305–314.

30. Hambury HJ, Watson J, Toole A, Sivyer A, Ashley DBJ. Interdisciplinary approaches in electrically mediated bone growth studies. Ann NY Acad Sci 1974;238:508–518.
31. Paterson DC, Carter RF, Tilbury RF, Ludbrook J, Savage JP. The effects of varying current levels of electrical stimulation. Clin Orthop Rel Res 1982;169:303–312.
32. Ohashi T. Electrical callus formation and its osteogenesis. J Jpn Orthop Ass 1982;56:615–633.
33. Inoue S, Ohashi T, Fukada E, Ashihara T. Electric stimulation of osteogenesis in the rat: Amperage of three different stimulation mehods. In: Brighton CT, Black J, Pollack S, editors. Electric Properties of Bone and Cartilage. Philadelphia: Grune and Stratton; 1979. p 199–213.
34. Marino AA, Becker RO. Electrical osteogenesis: an analysis (Let). Clin Orthop Rel Res 1977;123:280–282.
35. Bassett CAL, Pawluk RJ, Pilla AA. Acceleration of fracture repair by electromagnetic fields. A surgically noninvasive method. Ann NY Acad Sci 1974;238:242–262.
36. Werner FW, Spadaro JA. Engineering aspects of medical surgical instruments and devices. In: Barzeley ME, editor. Product Liability. New York: Matthew Bender Publ. Co; 1993.
37. Lavine LS, Lustrin I, Shamos MH, Rinaldi RA, Liboff AR. Electric enhancement of bone healing. Science 1972;175: 1118–1121.
38. Friedenberg ZB, Harlow MC, Brighton CT. Healing of nonunion in the medial malleolus by means of direct current: a case report. J Trauma 1971;11:883–885.
39. Friedenberg ZB, Roberts PG, Jr, Didizian NH, Brighton CT. Stimulation of fracture healing by direct current in the rabbit fibula. J Bone Joint Surg 1971;53A:1400–1408.
40. Luben RA, et al. Effects of electromagnetic stimuli on bone and bone cells in vitro inhibition of responses to parathyroid hormone by low-energy, low-frequency field. Proc Natl Acad Sci USA 1982;79:4180–4184.
41. Norton LA, Rodan GA, Bourret LA. Epiphyseal cartilage cAMP changes produced by electrical and mechanical perturbations. Clin Orthop Rel Res 1977;124:57.
42. Farndale RW, Murray JC. The action of pulsed magnetic fields on cyclic AMP levels in cultured fibroblasts. Biochim. Biphys Acta 1986;881:46–53.
43. Fitzsimmons RJ, Ryaby JT, Mohan S, Magee FP, Baylink DJ. Combined magnetic fields increase IGF-II in TE-85 human bone cell cultures. Endocrinology 1995;136:3100–3106.
44. Bodamyali T, Bhatt B, Hughes FJ, Winrow VR, Kanczler JM, Simon B, Abbott J, Blake DR, Stevens CR. Pulsing electromagnetic fields simultaneously induce osteogenesis and upregulate transcription of bone morphogentic proteins 2 and 4in rat osteoblasts in vitro. Biochem Biophys Res Commun 1998;250:458–461.
45. Zhuang H, Wang W, Seldes RM, Tahemia AD, Fan H, Brighton CT. Electrical stimulation induces the level of TGF-β1 mRNA in osteoblastic cells by a mechanism involving calcium/calmodulin pathway. Biochem Biophys Res Commun 1997;237:225–229.
46. Wang Q, Zhong S, Ouyang J, Jiang L, Zhang Z, Xie Y, Luo S. Osteogenesis of electrically stimulated bone cells mediated in part by calcium ions. Clin Orthop Rel Res 1996;348:259–268.
47. Brighton CT, McCluskey WP. Response of bone cells to a capacitively coupled electric field: inhibition of cyclic adenosine monophosphate response to parathroid hormone. J Orthop Res 1988;6:567–571.
48. Lorich DG, Brighton CT, Gupta R, Corsetti JR, Levine SE, Gelb ID, Seldes R, Pollack SR. Biochemical pathway mediating the response of bone cells to capacitive coupling. Clin Orthop Rel Res 1998;350:246–256.
49. Brighton CT, Pollack SR. Treatment of recalcitrant nonunion with a capacitively coupled electric field. J Bone Joint Surg 1985;67A:577–585.
50. De Haas WG, Morrison DM, Watson J. Non-invasive treatment of the tibia using electrical stimulation. J Bone Joint Surg 1980;62:465–470.
51. Watson J, Downes EM. Light-weight battery-operable orthopaedic stimulator for the treatment of long-bone nonunions using pulsed magnetic fields. Med Biol Eng Comput 1983;21: 509–510.
52. Bassett CAL, Pilla AA, Pawluk RJ. A non-operative salvage of surgically-resistant pseudarthroses non-unions by pulsing electromagnetic fields electromagnetic fields. Clinical Orthop Rel Res 1977;124:128–143.
53. Frykman GK, Taleisnik J, Peters G, Kaufman R, Helal B, Wood VE, Unsell RS. Treatment of nonunited scaphoid fractures by pulsed electromagnetic field and cast. J Hand Surg 1986;11:344–349.
54. Kort JS, Schink MM, Mitchell SN, Bassett CAL. Congenital pseudarthrosis of the tibia: Treatment with pulsing electromagnetic fields. Clin Orthop Rel Res 1982;165:124–136.
55. Bassett CAL, Schink-Ascani M. Long-term pulsed electromagnetic field (PEMF) results in congenital pseudarthrosis. Calc Tiss Int 1991;49:216–220.
56. Liboff AR. Geomagnetic cyclotron resonance in living cells. J Biol Phys 1985;13:99–102.
57. Bawin SM, Kazmarek KL, Adey WR. Effects of modulated VHF fields on the central nervous system. Ann NY Acad Sci 1975;247:74–81.
58. Blackman CF, Benane SG, Kinney LS, Joines WT, House DE. Effects of ELF fields on calcium-ion efflux from brain tissue in vivo. Rad Res 1982;92:510–520.
59. Liboff AR, Williams T, Jr., Strong DM, Wistar R, Jr. Time-varying magnetic fields: effect on DNA synthesis. Science 1984;223:818–820.
60. Takahashi K, Keneko I, Date M, Fukada E. Effect of pulsing electromagnetic fields on DNA synthesis in mammalian cells in culture. Experientia 1986;42:185–186.
61. Liboff AR. The charge-to-mass ICR signature in weak ELF bioelectromagnetic effects. In: Lin JC, editor. Advances in Electromagnetic Fields in Living Systems. Volume 4, New York: Kluwer; 2003.
62. Fitzsimmons RJ, Ryaby JT, Mohan JT, Magee FP, Baylink DJ. Combined magnetic fields increase insulin-like growth factor-II in Te-85 human osteosarcoma bone cell cultures. Endocrinology 1995;136:3100–3106.
63. Smith SD, Liboff AR, McLeod BR. Effects of resonant magnetic fields on chick femoral development in vitro. J Bioelect 1999;10:81–99.
64. Diebert MC, McLeod BR, Smith SD, Liboff AR. Ion resonance electromagnetic field stimulation of fracture healing in rabbits with a fibular ostectomy. J Orthop Res 1994;12:878–885.
65. McLeod BR, Liboff AR. Cyclotron resonance in cell membranes; The theory of the mechanism. In: Blank M, Findl E, editors. Mechanistic Approaches to Interactions of Electric Electromagnetic Fields with Living Systems. New York: Plenum Press; 1987; p 97–108.
66. Smith SD. personal Communication.
67. Rozek RJ, Sherman ML, Liboff AR, McLeod BR, Smith SD. Nifedipene is an antagonist to cyclotron resonance enhancement of 45Ca incorporation in human lymphocytes. Cell Calcium 1987;8:413–427.
68. Barker AT, Dixon RA, Sharrard WJW, Sutcliffe ML. Pulsed magnetic field therapy for tibial non-union. The Lancet 1984;1:994–996.
69. Sharrard WJW. A double-blind trial of pulsed electromagnetic field for delayed union of tibial fractures. J Bone Joint Surg Bri 1990;72B:347–355.

70. Mammi GI, Rocchi R, Cadossi R, Traina GC. Effect of PEMF on the healing of human tibial osteomoties: a double blind study. Clin Orthop 1993;288:246–253.
71. Scott G, King JB. A prospective double blind trial of electrical capacitive coupling in the treatment of non-union of long bones. J Bone Joint Surg 1994;76A:820–826.
72. Longo JA. The management of recalcitrant nonunions with combined magnetic fields. Orthop Trans 1998;22:408–409.
73. Mooney VA. randomized double blind prospective study of the efficacy of pulsed electromagnetic fields for interbody lumbar fusions. Spine 1990;15:708–715.
74. Goodwin CB, Brighton CT, Guyer RD, Guyer RD, Johnson JR, Light KI, Yuan HA. A double blind study of capacitively coupled electrical stimulation as an adjunct to lumbar spinal fusions. Spine 1999;24:1349–1357.
75. Linovitz RJ, Pathria M, Bernhardt M, Green D, Law MD, McGuire RA, Montesano P, Rechtine G, Salib RM, Ryaby JT, Faden JS, Ponder R, Muenz LR, Magee FP, Garfin SA. Combined magnetic fields accelerate increase spine fusion: a double-blind, randomized, placebo controlled study. Spine 2002;27:1383–1389.
76. Satter SA, Islam MA, Rabbani KS, Talukder MS. Pulsed electromagnetic fields for the treatment of bone fractures. Bangladesh Med Res Council Bull 1999;25:6–19.
77. Fredericks DC, Nepola JV, Baker JT, Abbott J, Simon B. Effects of pulsed electromagnetic fields on bone healing in a rabbit tibial osteomoty model. J Orthop Trauma 2000;14:93–100.
78. Lednev VV. Possible mechanism for the influence of weak magnetic fields on biological systems. Bioelectromagnetics 1991;12:71–75.
79. Blanchard JP, Blackman CF. Clarification amplification of an ion parametric resonance model for magnetic field interactions with biological systems. Bioelectromagnetics 1994;15: 217–238.
80. Prato FP, Kavaliers M, Thomas AW. Extremely low frequency magnetic fields can either increase or decrease analgesia in the land snail depending on field and light conditions. Bioelectromagnetics 2000;21:287–301.
81. Becker RO. personal communication.
82. Goodman R, Bassett CAL, Henderson AS. Pulsing electromagnetic fields induce cellular transcription. Science 1983;220:1283–1285.
83. Goodman R, Blank M. A non-thermal low-energy agent that induces stress response proteins: Magnetic fields. Cell Stress Chaperones 1998;3:79–88.

See also BONE AND TEETH, PROPERTIES OF; FUNCTIONAL ELECTRICAL STIMULATION; HUMAN SPINE, BIOMECHANICS OF.

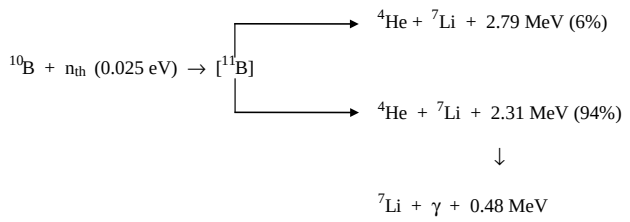
BORON NEUTRON CAPTURE THERAPY

THOMAS E BLUE
The Ohio State University
Columbus, Ohio

INTRODUCTION

High grade gliomas, and specifically glioblastoma multiforme (GBM), are still extremely resistant to all current forms of therapy, including surgery, chemotherapy, radiotherapy, immunotherapy, and gene therapy after decades of intensive research (1–5). Despite aggressive treatment using combinations of therapeutic modalities, the 5 year survival rate of patients diagnosed with GBM in the United States is less than a few percent (6,7). By the time they have had surgical resection of their tumors, malignant cells have infiltrated beyond the margins of resection and have spread into both gray and white matter (8,9). As a result, high grade supratentorial gliomas must be regarded as a whole brain disease (10). Glioma cells and their neoplastic precursors have biochemical properties that allow them to invade the unique extracellular environment of the brain (11,12) and biologic properties that allow them to evade a tumor associated host immune response (13). Chemo- and radiotherapy's inability to cure patients with high grade gliomas is due to their failure to eradicate microinvasive tumor cells within the brain. The challenge facing us is how to develop molecular strategies that can selectively target malignant cells with little or no effect on normal cells and tissues adjacent to the tumor. However, recent molecular genetic studies of glioma suggest that it may be much more complicated than this (14).

In theory, boron neutron capture therapy (BNCT) provides a way to selectively destroy malignant cells and spare normal cells. It is based on the nuclear capture and fission reactions that occur when boron-10, which is a nonradioactive constituent of natural elemental boron, is irradiated with low energy thermal neutrons to yield high linear energy-transfer (LET) alpha particles (^{4}He) and recoiling lithium-7 (^{7}Li) nuclei, as shown below.



In order for BNCT to be successful, a sufficient amount of ^{10}B must be selectively delivered to the tumor ($\sim$20 $\mu g \cdot g^{-1}$ weight or $\sim 10^9$ atoms/cell), and enough thermal neutrons must be absorbed by them to sustain a lethal $^{10}B(n, \alpha)\ ^7Li$ capture reaction. Since the high LET particles have limited boron pathlengths in tissue (5–9 μm), the destructive effects of these high energy particles is limited to cells containing boron. Clinical interest in BNCT has focused primarily on the treatment of high grade gliomas (15), and either cutaneous primaries (16) or cerebral metastases of melanoma (17), and most recently head and neck and liver cancer. Since BNCT is a biologically rather than physically targeted

ROLF F. BARTH
The Ohio State University
Columbus, Ohio
JEFFREY A. CODERRE
Massachusetts Institute of Technology
Cambridge, Massachusetts
M. GRAÇA
H. VICENTE
Louisiana State University
Baton Rouge, Louisiana

type of radiation treatment, the potential exists to destroy tumor cells dispersed in the normal tissue parenchyma, if sufficient amounts of ^{10}B and thermal neutrons are delivered to the target volume. This article covers radiobiological considerations upon which BNCT is based, boron agents and optimization of their delivery, neutron sources, which at this time are exclusively nuclear reactors, past and ongoing clinical studies, and critical issues that must be addressed if BNCT is to be successful. Readers interested in more in-depth coverage of these and other topics related to BNCT are referred to several recent reviews and monographs (15,18–20).

RADIOBIOLOGICAL CONSIDERATIONS

Types of Radiation Delivered

The radiation doses delivered to tumor and normal tissues during BNCT are due to energy deposition from three types of directly ionizing radiation that differ in their LET characteristics: (*1*) low LET γ rays, resulting primarily from the capture of thermal neutrons by normal tissue hydrogen atoms [$^{1}H(n,\gamma)^{2}H$]; (*2*) high LET protons, produced by the scattering of fast neutrons and from the capture of thermal neutrons by nitrogen atoms [$^{10}N(n,p)^{14}C$]; and (*3*) high LET, heavier charged alpha particles (stripped down ^{4}He nuclei) and lithium-7 ions, released as products of the thermal neutron capture and fission reactions with ^{10}B [$^{10}B(n,\alpha)^{7}Li$]. The greater density of ionizations along tracks of high LET particles results in an increased biological effect compared to the same physical dose of low LET radiation. Usually, this is referred to as relative biological effectiveness (RBE), which is the ratio of the absorbed dose of a reference source of radiation (e.g., X rays) to that of the test radiation that produces the same biological effect. Since both tumor and surrounding normal tissues are present in the radiation field, even with an ideal epithermal neutron beam, there will be an unavoidable, nonspecific background dose, consisting of both high and low LET radiation. However, a higher concentration of ^{10}B in the tumor will result in it receiving a higher total dose than that of adjacent normal tissues, which is the basis for the therapeutic gain in BNCT (21). As recently reviewed by one of us (18), the total radiation dose delivered to any tissue can be expressed in photon-equivalent units as the sum of each of the high LET dose components multiplied by weighting factors, which depend on the increased radiobiological effectiveness of each of these components.

Biological Effectiveness Factors

The dependence of the biological effect on the microdistribution of ^{10}B requires the use of a more appropriate term than RBE to define the biological effects of the $^{10}B(n,\alpha)^{7}Li$ reaction. Measured biological effectiveness factors for the components of the dose from this reaction have been termed compound biological effectiveness (CBE) factors and are drug dependent (21–23). The mode and route of drug administration, the boron distribution within the tumor, normal tissues, and even more specifically within cells, and even the size of the nucleus within the target cell population all can influence the experimental determination of the CBE factor. Therefore, CBE factors are fundamentally different from the classically defined RBE, which primarily is dependent on the quality (i.e., LET) of the radiation administered. The CBE factors are strongly influenced by the distribution of the specific boron delivery agent, and can differ substantially, although they all describe the combined effects of alpha particles and ^{7}Li ions. The CBE factors for the boron component of the dose are specific for both the boron-10 delivery agent and the tissue. A weighted gray (Gy) unit [Gy(w)] has been used to express the summation of all BNCT dose components and indicates that the appropriate RBE and CBE factors have been applied to the high LET dose components. However, for clinical BNCT the overall calculation of photon-equivalent [Gy(w)] doses requires a number of assumptions about RBEs, CBE factors, and the boron concentrations in various tissues that have been based on the currently available human or experimental data (24,25).

Clinical Dosimetry

The following biological weighting factors, summarized in Table 1, have been used in all of the recent clinical trials in patients with high grade glioma, using BPA in combination with an epithermal neutron beam. The $^{10}B(n,\alpha)^{7}Li$ component of the radiation dose to the scalp has been based on the measured boron concentration in the blood at the time of BNCT, assuming a blood:scalp boron concentration ratio of 1.5:1 (26,27,29) and a CBE factor for BPA in skin of 2.5 (29). An RBE of 3.2 has been used in all tissues for the high LET components of the beam: protons resulting from the capture reaction with nitrogen, and recoil protons resulting from the collision of fast neutrons with hydrogen (26,27,30). It must be emphasized that the tissue distribution of the boron delivery agent in humans should be similar to that in the experimental animal model in order to use the experimentally derived values for estimation of Gy(w) doses in clinical radiations.

Dose calculations become much more complicated when combinations of agents are used. At its simplest, this could be the two low molecular weight drugs boronophenylalanine (BPA) and sodium borocaptate (BSH). These have been shown to be highly effective when used in combination to treat F98 glioma bearing rats (31,32), and currently are being used in combination in a clinical study in Japan (33). Since it currently is impossible to know the true

Table 1. Assumptions Used in the Clinical Trials of BPA Based BNCT for Calculation of the $^{10}B(n,\alpha)^{7}Li$ Component of the Gy(w) Dose in Various Tissue

Tissue	Boron Concentration[a]	CBE Factor
Blood	measured directly	
Brain	1.0 × blood (26,27)	1.3 (23)
Scalp–skin	1.5 × blood (26–28)	2.5 (29)
Tumor	3.5 × blood (28)	3.8 (21)

[a] An RBE of 3.2 is used for the high LET component of the beam dose: protons from the $^{14}N(n,n)^{14}C$ reaction, and the recoil protons from fast neutron collisions with hydrogen. Literature references are given in parentheses.

biodistribution of each drug, dosimetric calculations in experimental animals have been based on independent boron determinations in other tumor bearing animals that have received the same doses of drugs but *not* BNCT. More recently, the radiation delivered has been expressed as a physical dose rather than using CBE factors to calculate an RBE equivalent dose (34). The calculations are further complicated if low and high molecular weight delivery agents are used in combination with one another. Tumor radiation dose calculations, therefore, are based on multiple assumptions regarding boron biodistribution, which may vary from patient to patient, as well as within different regions of the tumor and among tumor cells. However, normal brain boron concentrations are much more predictable and uniform, and therefore, it has been shown to be both *safe* and *reliable* to base dose calculations on normal brain tolerance.

BORON DELIVERY AGENTS

General Requirements

The development of boron delivery agents for BNCT began ~50 years ago and is an ongoing and difficult task of the highest priority. The most important requirements for a successful boron delivery agent are (*1*) low systemic toxicity and normal tissue uptake with high tumor uptake and concomitantly high tumor/brain (T/Br) and tumor/blood (T/Bl) concentration ratios (>3-4:1); (*2*) tumor concentrations in the range of ~20 μg $^{10}B \cdot g^{-1}$ tumor; (*3*) rapid clearance from blood and normal tissues and persistence in tumor during BNCT. However, at this time *no* single boron delivery agent fulfills all of these criteria. With the development of new chemical synthetic techniques and increased knowledge of the biological and biochemical requirements needed for an effective agent and their modes of delivery, a number of promising new boron agents has emerged (see examples in Fig. 1). The major challenge in their development has been the requirement for selective tumor targeting in order to achieve boron concentrations sufficient to deliver therapeutic doses of radiation to the tumor with minimal normal tissue toxicity. The selective destruction of GBM cells in the presence of normal cells represents an even greater challenge compared to malignancies at other anatomic sites, since high grade gliomas are highly infiltrative of normal brain, histologically complex, and heterogeneous in their cellular composition.

First- and Second-Generation Boron Delivery Agents

The clinical trials of BNCT in the 1950s and early 1960s used boric acid and some of its derivatives as delivery agents, but these simple chemical compounds were nonselective, had poor tumor retention, and attained low T/Br ratios (35,36). In the 1960s, two other boron compounds emerged from investigations of hundreds of low molecular weight boron-containing chemicals, one, (L)-4-dihydroxyborylphenylalanine, referred to as BPA (compound **1**) was based on arylboronic acids (37), and the other was based on a newly discovered polyhedral borane anion, sodium mercaptoundecahydro-*closo*-dodecaborate (38), referred to as BSH (compound **2**). These "second" generation compounds had low toxicity, persisted longer in animal tumors compared with related molecules, and their T/Br and T/Bl boron ratios were > 1. As described later in this article, ^{10}B enriched BPA, complexed with fructose to improve its water solubility, and BSH have been used clinically in Japan, the United States, and Europe. Although these drugs are not ideal, their safety following intravenous (i.v.) administration has been established. Over the past 20 years, several other classes of boron-containing compounds have been designed and synthesized in order to fulfill the requirements indicated at the beginning of this section. Detailed reviews of the state-of-the-art in compound development for BNCT have been published (39–42), and in this overview, only the main classes of compounds are summarized with an emphasis on recently published work in the area. The general biochemical requirements for an effective boron delivery agent are also discussed.

Third Generation Boron Delivery Agents

So-called "third" generation compounds mainly consist of a stable boron group or cluster attached via a hydrolytically stable linkage to a tumor-targeting moiety, such as low molecular weight biomolecules or monoclonal antibodies (MoAbs). For example, the targeting of the epidermal growth factor receptor (EGFR) and its mutant isoform EGFRvIII, which are overexpressed in gliomas and squamous cell carcinomas of the head and neck, also has been one such approach (43). Usually, these low molecular weight biomolecules have been shown to have selective targeting properties and many are at various stages of development for cancer chemotherapy, photodynamic therapy (PDT) or antiviral therapy. The tumor cell nucleus and DNA are especially attractive targets since the amount of boron required to produce a lethal effect may be substantially reduced, if it is localized within or near the nucleus (44). Other potential subcellular targets are mitochondria, lysosomes, endoplasmic reticulum, and the Golgi apparatus. Water solubility is an important factor for a boron agent that is to be administered systemically, while lipophilicity is necessary for it to cross the blood–brain barrier (BBB) and diffuse within the brain and the tumor. Therefore, amphiphilic compounds possessing a suitable balance between hydrophilicity and lipophilicity have been of primary interest since they should provide the most favorable differential boron concentrations between tumor and normal brain, thereby enhancing tumor specificity. However, for low molecular weight molecules that target specific biological transport systems and/or are incorporated into a delivery vehicle (e.g., liposomes) the amphiphilic character is not as crucial. The molecular weight of the boron-containing delivery agent also is an important factor, since it determines the rate of diffusion both within the brain and the tumor.

LOW MOLECULAR WEIGHT AGENTS

Boron-Containing Amino Acids and Polyhedral Boranes

Recognizing that BPA and BSH are not ideal boron delivery agents, considerable effort has been directed toward

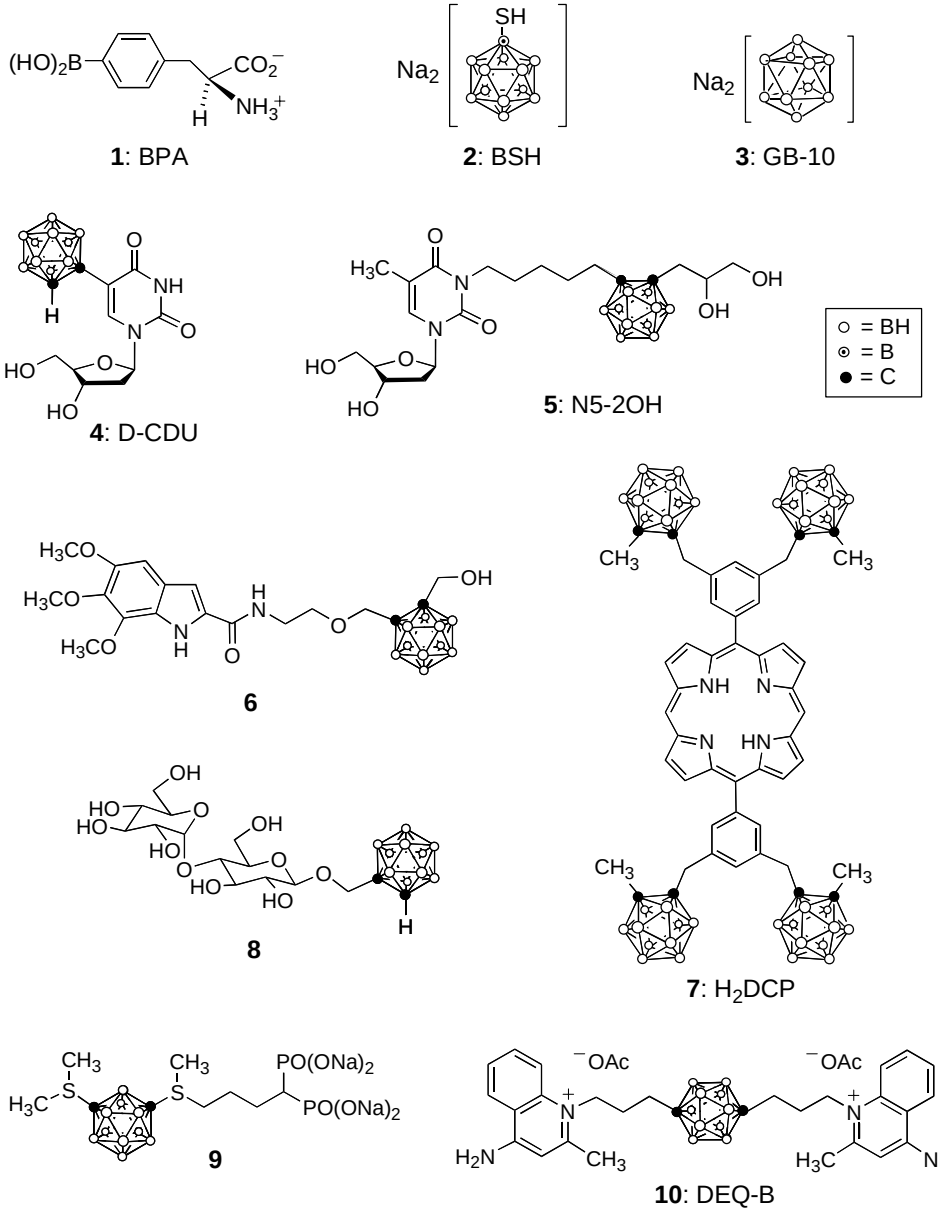


Figure 1. Some low molecular weight BNCT agents under investigation. Compound **1** (BPA) and compound **2** (BSH) are currently in clinical use in the United States, Japan, and Europe. Compound **3** (GB-10) has shown promise in animal models, as have the nucleoside derivatives D-CDU (compound **4**) and N5-2OH (compound **5**). Compound **6**, a trimethoxyindole derivative, has shown promise *in vitro* and compound **7**, a porphyrin derivative, was shown to be tumor selective. The maltose derivative **8** has shown low cytotoxicity and tumor cell uptake *in vitro*, the biphosphonate **9** has tumor targeting ability and the dequalinium derivative DEQ-B (compound **10**) has shown promise in *in vitro* studies.

the design and synthesis of third generation, boron-containing amino acids and functionalized polyhedral borane clusters. Examples include various derivatives of BPA and other boron-containing amino acids (e.g., glycine, alanine, aspartic acid, tyrosine, cysteine, methionine), as well as non-naturally occurring amino acids (45–50). The most recently reported delivery agents contain one or more boron clusters and concomitantly larger amounts of boron by weight compared with BPA. The advantages of such compounds are that they potentially can deliver higher concentrations of boron to tumors without increased toxicity. The polyhedral borane dianions, *closo*-$B_{10}H_{10}^{2-}$ and *closo*-$B_{12}H_{12}^{2-}$ and the icosahedral carboranes *closo*-$C_2B_{10}H_{12}$ and *nido*-$C_2B_9H_{12}^{-}$, have been the most attractive boron clusters for linkage to targeting moieties, due to their relatively easy incorporation into organic molecules, high boron content, chemical and hydrolytic stability, hydrophobic character and, in most cases, their negative charge. The simple sodium salt of *closo*-$B_{10}H_{10}^{2-}$ (GB-10, compound **3**) has been shown to have tumor-targeting ability and low systemic toxicity in animal models (42) and has been considered as a candidate for clinical evaluation (51). Other polyhedral borane anions with high boron content include derivatives of $B_{20}H_{18}^{2-}$, although these compounds have shown little tumor specificity, and therefore may be better candidates for encapsulation into either targeted or non-targeted liposomes (52,53) and folate receptor targeting, boron containing polyamidoamino (PAMAM) dendrimers (54) and liposomes (55). Boron-containing dipeptides also have shown low toxicity and good tumor-localizing properties (56,57).

Biochemical Precursors and DNA Binding Agents

Several boron-containing analogs of the biochemical precursors of nucleic acids, including purines, pyrimidines,

nucleosides, and nucleotides, have been synthesized and evaluated in cellular and animal studies (58–62). Some of these compounds [e.g., β-5-*o*-carboranyl-2'-deoxyuridine (D-CDU, compound **4**] and the 3-(dihydroxypropyl-carboranyl-pentyl)thymidine derivative N5-2OH (compound **5**), have shown low toxicities, selective tumor cell uptake, and significant rates of phosphorylation into the corresponding nucleotides (63–65). Intracellular nucleotide formation potentially can lead to enhanced tumor uptake and retention of these types of compounds (64,65).

Another class of low molecular weight delivery agents are boron-containing DNA binding molecules (e.g., alkylating agents, intercalators, groove binders, and polyamines). Some examples are derivatives of aziridines, acridines, phenanthridines (compound **6**), trimethoxyindoles, carboranylpolyamines, Pt(II)–amine complexes, di- and tribenzimidazoles (66–69). A limitation of boron-containing polyamines is their frequently observed *in vitro* and *in vivo* toxicity, although promising derivatives with low cytotoxicity have been synthesized (70–73). Other nuclear-targeting molecules are *nido*-carboranyl oligomeric phosphate diesters (OPDs). Despite their multiple negative charges, OPDs have been shown to target the nuclei of TC7 cells following microinjection (74), suggesting that the combination of OPDs with a cell-targeting molecule capable of crossing the plasma membrane could provide both selectivity and nuclear binding. Such a conjugate has been designed and synthesized (75), although its biological evaluation has yet to be reported.

Boron-Containing Porphyrins and Related Structures

Several boron-containing fluorescent dyes, including porphyrin, tetrabenzoporphyrin, and phthalocyanine derivatives have been synthesized and evaluated (76–79). These have the advantage of being easily detected and quantified by fluorescence microscopy, and have the potential for interacting with DNA due to their planar aromatic structures. Among these macrocycles, boron-containing porphyrins (e.g., compound **7**: H_2DCP) have attracted special attention due to their low systemic toxicity compared with other dyes, easy synthesis with high boron content, and their remarkable stability (79–82). Porphyrin derivatives have been synthesized that contain up to 44% boron by weight using *closo*- or *nido*-carborane clusters linked to the porphyrin macrocycle via ester, amide, ether, methylene, or aromatic linkages (76–85). The nature of these linkages is believed to influence their stability and systemic toxicity. Therefore, with these and other boron delivery agents, chemically stable carbon–carbon linkages have been preferred over ester and amide linkages that potentially can be cleaved *in vivo*. Boron-containing porphyrins have excellent tumor-localizing properties (76–82) and have been proposed for dual application as boron delivery agents and photosensitizers for PDT of brain tumors (85–91). Our own preliminary data with H_2TCP (tetra[*nido*-carboranylphenyl]porphyrin), administered intracerebrally by means of convection enhanced delivery (CED) to F98 glioma bearing rats, showed tumor boron concentrations of 150 $\mu g \cdot g^{-1}$ tumor with concomitantly low normal brain and blood concentrations (92). Ozawa et al. recently described a newly synthesized polyboronated porphyrin, designated TABP-1, which was administered by CED to nude rats bearing intracerebral implants of the human glioblastoma cell line U-87 MG (93). High tumor and low blood boron concentrations were observed and both we and Ozawa have concluded that direct intracerebral administration of the carboranyl porphyrins by CED is superior to systemic administration. Furthermore, despite the bulkiness of the carborane cages, carboranylporphyrins have been shown to interact with DNA and thereby produce *in vitro* DNA damage following light activation (94,95). Boronated phthalocyanines have been synthesized, although these compounds usually have had decreased water solubility and an increased tendency to aggregate compared to the corresponding porphyrins (76,77,86,87). Boron-containing acridine molecules also have been reported to selectively deliver boron to tumors with high T/Br and T/Bl ratios, whereas phenanthridine derivatives were found to have poor specificity for tumor cells (94–98).

Other Low Molecular Weight Boron Delivery Agents

Carbohydrate derivatives of BSH and other boron-containing glucose, mannose, ribose, gulose, fucose, galactose, maltose (e.g., compound **8**) and lactose molecules have been synthesized, and some of these compounds have been evaluated in both *in vitro* and *in vivo* studies (99–105). These compounds usually are highly water soluble and as a possible consequence of this, they have shown both low toxicity and uptake in tumor cells. It has been suggested that these hydrophilic low molecular weight derivatives have poor ability to cross tumor cell membranes. However, they might selectively accumulate within the glycerophospholipid membrane bilayer and in other areas of the tumor, such as the vasculature.

Low molecular weight boron-containing receptor-binding molecules have been designed and synthesized. These have been mainly steroid hormone antagonists, such as derivatives of tamoxifen, 17β-estradiol, cholesterol, and retinoic acid (106–110). The biological properties of these agents depend on the density of the targeted receptor sites, although to date very little biological data have been reported. Other low molecular weight boron-containing compounds that have been synthesized include phosphates, phosphonates (e.g., compound **9**) phenylureas, thioureas, nitroimidazoles, amines, benzamides, isocyanates, nicotinamides, azulenes, and dequalinium derivatives (e.g., dequalinium-B, compound **10**) (111–113). Since no single chemical compound, as yet synthesized, has the requisite properties, the use of *multiple* boron delivery agents is probably essential for targeting different subpopulations of tumor cells and subcellular sites. Furthermore, lower doses of each individual agent would be needed, which could reduce systemic toxicity while at the same time enhancing tumor boron levels to achieve a therapeutic effect.

HIGH MOLECULAR WEIGHT AGENTS

Monoclonal Antibodies, Other Receptors Targeting Agents and Liposomes

High molecular weight delivery agents (e.g., MoAbs and their fragments), which can recognize a tumor-associated epitope, have been (114–116) and continue to be of interest to us (117,118) as boron delivery agents. Although they can be highly specific, only very small quantities reach the brain and tumor following systemic administration (119) due to their rapid clearance by the reticuloendothelial system and the BBB, which effectively limits their ability to cross capillary vascular endothelial cells. Boron-containing bioconjugates of epidermal growth factor (EGF) (120,121), the receptor which is overexpressed on a variety of tumors, including GBM (122,123), also have been investigated as potential delivery agents to target brain tumors. However, it is unlikely that either boronated antibodies or other bioconjugates would attain sufficiently high concentrations in the brain following systemic administration, but, as described later in this section, direct intracerebral delivery could solve this problem. Another approach would be to directly target the vascular endothelium of brain tumors using either boronated MoAbs or VEGF, which would recognize amplified VEGF receptors. The use of boron-containing VEGF bioconjugates would obviate the problem of passage of a high molecular weight agent across the BBB, but their use would most likely require repeated applications of BNCT, since tumor neovasculature can continuously regenerate. Backer et al. reported that targeting a Shiga-like toxin-VEGF fusion protein was selectively toxic to vascular endothelial cells overexpressing VEGFR-2 (124). Recently, a bioconjugate has been produced by chemically linking a heavily boronated PAMAM dendrimer to VEGF (125). This selectively targeted tumor blood vessels overexpressing VEGFR-2 in mice bearing 4T1 breast carcinoma. There also has been a longstanding interest on the use of boron-containing liposomes as delivery agents (52,53,126,127), but their size has limited their usefulness as brain tumor targeting agents, since they are incapable of traversing the BBB unless they have diameters <50 nm (128). If, on the other hand, they were administered intracerebrally or were linked to an actively transported carrier molecule (e.g., transferin), or alternatively if the BBB was transiently opened, these could be very useful delivery agents, especially for extracranial tumors (e.g., liver cancer).

Recent work of one of us (R.F.B.) has focused on the use of a chemeric MoAb, cetuximab (IMC-C225 also known as Erbitux), produced by ImClone Systems, Inc. This antibody recognizes both wild-type EGFR and its mutant isoform, EGFRvIII (129), and has been approved for clinical use by the U.S. Food and Drug Administration (FDA) for the treatment of EGFR(+) recurrent colon cancer. Using previously developed methodology (114), a precision macromolecule, a polyamido amino (PAMAM or "starburst") dendrimer has been heavily boronated and then linked by means of heterobifunctional reagents to EGF (121), cetuximab (118) or another MoAb, L8A4, which is specifically directed against EGFRvIII (130). In order to completely bypass the BBB, the bioconjugates were administered by either direct intratumoral (i.t.) injection (131) or CED (132) to rats bearing intracerebral implants of the F98 glioma that had been genetically engineered to express either wildtype EGFR (131) or EGFRvIII (133). Administration by either of these methods resulted in tumor boron concentrations that were in the therapeutic range (i.e., $\sim 20\,\mu g \cdot g^{-1}$ wt^{-1} tumor). Similar data also were obtained using boronated EGF, and based on the favorable uptake of these bioconjugates, therapy studies were initiated at the Massachusetts Institute of Technology nuclear reactor (MITR). The mean survival times (MST) of animals that received either boronated cetuximab (134) or EGF (135) were significantly prolonged compared to those of animals bearing receptor negative tumors. A further improvement in MSTs was seen if the animals received BPA, administered i.v., in combination with the boronated bioconjugates, thereby validating our thesis that combinations of agents may be superior to any single agent (32). As can be seen from the preceding discussion, the design and synthesis of low and high molecular weight boron agents have been the subject of intensive investigation. However, optimization of their delivery has not received enough attention, but nevertheless is of critical importance.

OPTIMIZING DELIVERY OF BORON CONTAINING AGENTS

General Considerations

Delivery of boron agents to brain tumors is dependent on (*1*) the plasma concentration profile of the drug, which depends on the amount and route of administration; (*2*) the ability of the agent to traverse the BBB; (*3*) blood flow within the tumor, and (*4*) the lipophilicity of the drug. In general, a high steady-state blood concentration will maximize brain uptake, while rapid clearance will reduce it, except in the case of intraarterial (i.a.) drug administration. Although the i.v. route currently is being used clinically to administer both BSH and BPA, this may not be ideal and other strategies may be needed to improve their delivery. Delivery of boron-containing drugs to extracranial tumors, such as head and neck and liver cancer, present a different set of problems, including nonspecific uptake and retention in adjacent normal tissues.

Intra-arterial Administration with or without Blood–Brain Barrier Disruption

As shown in experimental animal studies (31,32,134–136) Enhancing the delivery of BPA and BSH can have a dramatic effect both on increasing tumor boron uptake and the efficacy of BNCT. This has been demonstrated in the F98 rat glioma model where intracarotid (i.c.) injection of either BPA or BSH doubled the tumor boron uptake compared to that obtained by i.v. injection (31). This was increased fourfold by disrupting the BBB by infusing a hyperosmotic (25%) solution of mannitol via the internal carotid artery. Mean survival times (MST)of animals that received either BPA or BSH i.c. with BBB-D were increased 295 and 117%, respectively, compared to irradiated controls (31). The best survival data were obtained using both BPA and BSH in

combination, administered by i.c. injection with BBB-D. The MST was 140 days with a cure rate of 25%, compared to 41 days following i.v. injection with no long-term surviving animals (32). Similar data have been obtained using a rat model for melanoma metastatic to the brain. BPA was administered i.c. to nude rats bearing intracerebral implants of the human MRA 27 melanoma with or without BBB-D. The MSTs were 104–115 days with 30% long-term survivors compared to a MST of 42 days following i.v. administration (134). A similar enhancement in tumor boron uptake and survival was observed in F98 glioma bearing rats following i.c. infusion of the bradykinin agonist, RMP-7 (receptor mediated permeabilizer-7), now called Cereport (136,137). In contrast to the increased tumor uptake, normal brain boron values at 2.5 h following i.c. injection were very similar for the i.v. and i.c. routes with or without BBB-D. Since BNCT is a binary system, normal brain boron levels only are of significance at the time of irradiation and high values at earlier time points are inconsequential. These studies have shown that a significant therapeutic gain can be achieved by optimizing boron drug delivery, and this should be important for both ongoing and future clinical trials using BPA and/or BSH.

Direct Intracerebral Delivery

Different strategies may be required for other low molecular weight boron-containing compounds whose uptake is cell cycle dependent, such as boron-containing nucleosides, where continuous administration over a period of days may be required. We recently have reported that direct i.t. injection or CED of the borononucleoside N5-2OH (compound **5**) were both effective in selectively delivering potentially therapeutic amounts of boron to rats bearing intracerebral implants of the F98 glioma (61). Direct i.t. injection or CED most likely will be necessary for a variety of high molecular weight delivery agents such as boronated MoAbs (138) and ligands such as EGF (132), as well as for low molecular weight agents (e.g., nucleosides and porphyrins). Recent studies have shown that CED of a boronated porphyrin derivative similar to compound **7**, designated H_2DCP, resulted in the highest tumor boron values and T/Br and T/Bl ratios that have been seen with any of the boron agents that have been studied (92).

NEUTRON SOURCES FOR BNCT

Nuclear Reactors

Neutron sources for BNCT currently are limited to nuclear reactors and in the present section only information that is described in more detail in a recently published review will be summarized (139). Reactor derived neutrons are classified according to their energies as thermal E_n <0.5 eV), epithermal (0.5 eV<E_n<10 keV), or fast (E_n>10 keV). Thermal neutrons are the most important for BNCT since they usually initiate the $^{10}B(n,\alpha)^7Li$ capture reaction. However, because they have a limited depth of penetration, epithermal neutrons, which lose energy and fall into the thermal range as they penetrate tissues, are now preferred for clinical therapy. A number of reactors with very good neutron beam quality have been developed and currently are being used clinically. These include (*1*) MITR, shown schematically in Fig. 2 (140); (*2*) clinical reactor at Studsvik Medical AB in Sweden (141); (*3*) the FRi1 clinical reactor in Helsinki, Finland (142); (*4*) R2-0 High Flux Reactor (HFR) at Petten in the Netherlands (143); (*5*) LVR-15 reactor at the Nuclear Research Institute (NRI) in Rez, Czech Republic (144); (*6*) Kyoto University Research Reactor (KURR) in Kumatori, Japan (145); (*7*) JRR4 at the Japan Atomic Energy Research Institute (JAERI) (146); and (*8*) the RA-6 CNEA reactor in Bariloche, Argentina (147). Other reactor facilities are being designed, notably the TAPIRO reactor at the ENEA Casaccia Center near Rome, Italy, which is unique in that it will be a low-power fast-flux reactor (148), and a facility in South Korea. Two reactors that have been used in the past for clinical BNCT are the Musashi Institute of Technology (MuITR) reactor in Japan and the Brookhaven Medical Research Reactor (BMRR) at the Brookhaven National Laboratory (BNL) in Upton, Long Island, New York (26,27,149). The MuITR was used by Hatanaka (150) and later by Hatanaka and Nakagawa (151). The BMRR was used for the clinical trial that was conducted at the Brookhaven National Laboratory between 1994 and 1999 (27,152) and the results are described in detail later in this section. Due to a variety of reasons, including the cost of maintaining the BMRR, it has been deactivated and is no longer available for use.

Reactor Modifications

Two approaches are being used to modify reactors for BNCT. The first or direct approach, is to moderate and filter neutrons that are produced in the reactor core. The second, the fission converter-plate approach, is indirect in that neutrons from the reactor core create fissions within a converter-plate that is adjacent to the moderator assembly, and these produce a neutron beam at the patient port. The MITR (153), which utilizes a fission converter-plate, currently sets the standard for the world for the combination of high neutron beam quality and short treatment time. It operates at a power of 5 MW and has been used for clinical as well as experimental studies for BNCT. Although the power is high compared to the majority of other reactors that are being used, the treatment time is unusually short, since it utilizes a fission converter-plate to create the neutron beam. All other reactors use the direct approach to produce neutron beams for BNCT. Three examples are the FiR1 reactor in Finland (142), the Studsvik reactor in Sweden (141), and the Washington State University (WSU) reactor in the United States (154), which was built for the treatment of both small and large experimental animals.

Accelerators

Accelerators also can be used to produce epithermal neutrons and accelerator based neutron sources (ABNSs) are being developed in a number of countries (155–161), and interested readers are referred to a recently published detailed review on this subject (28). For ABNSs, one of the more promising nuclear reactions involves bombarding a ^7Li target with 2.5 MeV protons. The average energy of

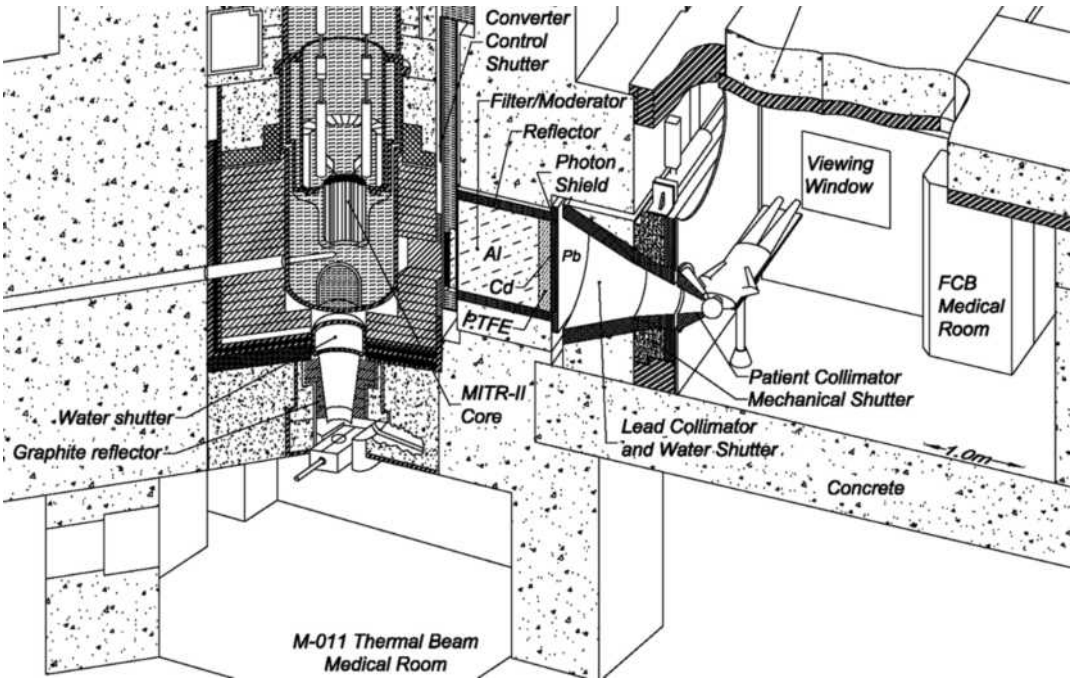


Figure 2. Schematic diagram of the MITR. The fission converter based epithermal neutron irradiation (FCB) facility is housed in the experimental hall of the MITR and operates in parallel with other user applications. The FCB contains an array of 10 spent MITR-II fuel elements cooled by forced convection of heavy water coolant. A shielded horizontal beam line contains an aluminum and Teflon filter-moderator to tailor the neutron energy spectrum into the desired epithermal energy range. A patient collimator defines the beam aperture and extends into the shielded medical room to provide circular apertures ranging from 16 to 8 cm in diameter. The in-air epithermal flux for the available field sizes ranges from 3.2 to 4.6×10^9 n·cm^{-2}·s^{-1} at the patient position. The measured specific absorbed doses are constant for all field sizes and are well below the inherent background of 2.8×10^{-12} RBE Gy·cm^2·n^{-1} produced by epithermal neutrons in tissue. The dose distributions achieved with the FCB approach the theoretical optimum for BNCT.

the neutrons that are produced is 0.4 MeV and the maximum energy is 0.8 MeV. Reactor derived fission neutrons have greater average and maximum energies than those resulting from the ^{7}Li(p,n)^{7}Be reaction. Consequently, the thickness of the moderator material that is necessary to reduce the energy of the neutrons from the fast to the epithermal range is less for an ABNS than it is for a reactor. This is important since the probability that a neutron will be successfully transported from the entrance of the moderator assembly to the treatment port decreases as the moderator assembly thickness increases. Due to lower and less widely distributed neutron source energies, ABNS potentially can produce neutron beams with an energy distribution that is equal to or better than that of a reactor. However, reactor derived neutrons can be well collimated, while on the other hand, it may not be possible to achieve good collimation of ABNS neutrons at reasonable proton beam currents. The necessity of good collimation for the effective treatment of GBM, is an important and unresolved issue that may affect usefulness of ABNS for BNCT. The ABNSs are also compact enough to be sited in hospitals thereby allowing for more effective, but technically more complicated procedures to carry out BNCT. However, to date, no accelerator has been constructed with a beam quality comparable to that of the MITR, which can be sited in a hospital and that provides a current of sufficient magnitude to treat patients in $<$30 min. Furthermore, issues relating to target manufacture and cooling must be solved before ABNS become a reality. The ABNS that is being developed at the University of Birmingham in England, by modifying a Dynamitron linear electrostatic accelerator (155), may be the first facility where patients will be treated, although progress has been slow. Another ABNS being constructed by LINAC Systems, Inc. in Albuquerque, New Mexico (162), and this could be easily sited in a hospital and produce an epithermal neutron beam.

Beam Optimization

For both reactors and ABNSs, a moderator assembly is necessary to reduce the energy of the neutrons to the epithermal range. The neutrons comprising the neutron beam have a distribution of energies and are accompanied by unwanted X rays and gamma photons. A basic tenet of BNCT is that the dose of neutrons delivered to the target volume should not exceed the tolerance of normal tissues, and this applies to neutron beam design, as well as to treatment planning (25). The implications of this for beam design is that the negative consequences of increased normal tissue damage for a more energetic neutron beams at shallow depths, outweighs the benefits of more deeply penetrating energetic neutrons. For fission reactors, the

average energy of the neutrons produced is ~2 MeV, but small numbers have energies as high as 10 MeV. There is generally a trade off between treatment time and the optimum beam for patient treatment in terms of the energy distribution of the neutrons and the contamination of the neutron beam with X rays and gamma photons. Not surprisingly, reactors with the shortest treatment time (i.e., the highest normal tissue dose rate) operate at the highest power, since the number of neutrons that is produced per unit time is proportional to the power, measured in megawatts. Furthermore, high beam quality is most easily achieved using reactors with high power, since a larger fraction of the neutrons can be filtered, as the neutrons traverse the moderator assembly without making the treatment time exceedingly long.

CLINICAL STUDIES OF BNCT FOR BRAIN TUMORS

Early Trials

Although the clinical potential of BNCT was recognized in the 1930s (163), it was not until the 1950s that the first clinical trials were initiated by Farr at the BNL (145,163) and by Sweet and Brownell at the Massachusetts General Hospital (MGH) using the MIT reactor (36,164,165). The disappointing outcomes of these trials, which ended in 1961 and subsequently were carefully analyzed by Slatkin (166), were primarily attributable to (*1*) inadequate tumor specificity of the inorganic boron chemicals that had been used as capture agents; (*2*) insufficient tissue penetrating properties of the thermal neutron beams; and (*3*) high blood boron concentrations that resulted in excessive damage to normal brain vasculature and to the scalp (36,164,165).

Japanese Clinical Trials

Clinical studies were resumed by Hatanaka in Japan in 1967, following a 2 year fellowship in Sweet's laboratory at the MGH, using a thermal neutron beam and BSH, which had been developed as a boron delivery agent by Soloway at the MGH (38). In Hatanaka's procedure (150,151), as much of the tumor was surgically removed as possible (debulking), and at some time thereafter, BSH (compound **2**) was administered by a slow infusion, usually intra-arterially (150), but later intravenously (151). Later (12–14 h) BNCT, was carried out at one or another of several different nuclear reactors. Since thermal neutrons have a limited depth of penetration in tissue, this necessitated reflecting the skin and raising the bone flap in order to directly irradiate the exposed brain. This eliminated radiation damage to the scalp and permitted treatment of more deep-seated residual tumors. As the procedure evolved over time, a ping–pong ball or silastic sphere was inserted into the resection cavity as a void space to improve neutron penetration into deeper regions of the tumor bed and adjacent brain (150,151,167,168). This is a major difference between the procedure carried out by Hatanaka, Nakagawa and other Japanese neurosurgeons and the BNCT protocols that have been carried out in the United States and Europe, which have utilized epithermal neutron beams that have not required reflecting the scalp and raising the bone flap at the time of irradiation. This has made it difficult to directly compare the Japanese clinical results with those obtained elsewhere, and this has continued on until very recently when the Japanese started using epithermal neutron beams (33). Most recently, Miyatake et al. initiated a clinical study utilizing the combination of BSH and BPA, both of which were administered i.v. at 12 and 1 h, respectively, prior to irradiation with an epithermal neutron beam (33). A series of 11 patients with high grade gliomas have been treated, and irrespective of the initial tumor volume, magnetic resonance imaging (MRI) and computed tomography (CT) images showed a 17–51% reduction in tumor volume that reached a maximum of 30–88%. However, the survival times of these patients were not improved over historical controls and further studies are planned to improve the delivery of BPA and BSH, which may enhance survival.

Analysis of the Japanese Clinical Results

Retrospective analysis of subgroups of patients treated in Japan by Hatanaka and Nakagawa (167,168) have described 2, 5, and 10 year survival rates (11.4, 10.4, and 5.7%, respectively) that were significantly better than those observed among patients treated with conventional, fractionated, external beam photon therapy. However, a cautionary note was sounded by Laramore and Spence (169) who analyzed the survival data of a subset of 12 patients from the United States who had been treated by Hatanaka between 1987 and 1994. They concluded that there were no differences in their survival times compared to those of age matched controls, analyzed according to the stratification criteria utilized by Curran et al. (6). In a recent review of Hatanaka's clinical studies, Nakagawa reported that the physical dose from the $^{10}B(n,\alpha)^{7}Li$ reaction, delivered to a target point 2 cm beyond the surgical margin, correlated with survival (168). For 66 patients with GBMs, those who survived <3 years ($n = 60$) had a minimum target point dose of 9.5 ± 5.9 Gy, whereas those who survived >3 years ($n = 6$) had a minimum target point dose of 15.6 ± 3.1 Gy from the $^{10}B(n,\alpha)^{7}Li$ reaction (168). The boron concentrations in brain tissue at the target point, which are required to calculate the physical radiation dose attributable to the $^{10}B(n,\alpha)^{7}Li$ capture reaction, were estimated to be 1.2X that of the patient's blood boron concentration (170).

OTHER RECENT AND ONGOING CLINICAL TRIALS

Beginning in 1994 a number of clinical trials, summarized in Table 2, were initiated in the United States and Europe. These marked a transition from low energy thermal neutron irradiation to the use of higher energy epithermal neutron beams with improved tissue penetrating properties, which obviated the need to reflect skin and bone flaps prior to irradiation. Up until recently, the procedure carried out in Japan required neurosurgical intervention immediately prior to irradiation, whereas the current epithermal neutron-based clinical protocols are radiotherapeutic procedures, performed several weeks after debulking surgery. Clinical trials for patients with brain tumors were initiated at a number of locations including (*1*) the

Table 2. Summary of Current or Recently Completed Clinical Trials of BNCT for the Treatment of Glioblastoma

Facility	Number of Patients	Duration of Administration	Drug	Dose, $mg \cdot kg^{-1}$	Boron Conc.,[a] $\mu g^{10}B \cdot g^{-1}$	Estimated Peak Normal Brain Dose, Gy(w)	Average Normal Brain Dose, Gy(w)	References
HTR, MuITR, JRR, KURR, Japan	>250 (1968–present)	1 h 1 h	BSH BPA	100 250	~20–30	13 Gy-Eq[b] ^{10}B component	Nd	168,169 33
HFR, Petten, The Netherlands	26 (1997–)	$100\ mg \cdot kg^{-1} \cdot min^{-1}$	BSH	100	30[c]	8.6–11.4 Gy-Eq[d] ^{10}B component	Nd	177
LVR-15, Rez, Czech Republic	5 (2001–present)	1 h	BSH	100	~20–30	<14.2	<2	178
BMRR Brookhaven	53 (1994–1999)	2 h	BPA	250–330	12–16	8.4–14.8	1.8–8.5	153,179
MITR-II, M67 MIT	20[e](1996–1999)	1–1.5 h	BPA	250–350	10–12	8.7–16.4	3.0–7.4	176
MITR-II, FCB MIT	6 (2001–2003)	1.5 h	BPA	350	~15			Unpublished
Studsvik AB Sweden	17 (30)[f](2001–2005)	6 h	BPA	900	24 (range: 15–34)	7.3–15.5	3.6–6.1	142
Fir I, Helsinki Finland	18 (1999–present) protocol P-01	2 h	BPA	290–400	12–15	8–13.5	3–6<7	143
Fir 1, Helsinki Finland	3 (2001–present)[g] protocol P-03	2 h	BPA	290	12–15	<8	2–3 <6	143

[a] During the irradiation.

[b] ^{10}B physical dose component dose to a point 2 cm deeper than the air-filled tumor cavity.

[c] Four fractions, each with a BSH infusion, $100\ mg \cdot kg^{-1}$ the first day, enough to keep the average blood concentration at $30\ \mu g\ ^{10}B \cdot g^{-1}$ during treatment on days 2–4.

[d] ^{10}B physical dose component at the depth of the thermal neutron fluence maximum.

[e] Includes two intracranial melanomas.

[f] J. Capala, unpublished, personal communication.

[g] Retreatment protocol for recurrent glioblastoma.

BMR at BNL from 1994–1999 for GBM using BPA with one or two neutron radiations, given on consecutive days (171–173); (*2*) the MITR from 1996–1999 for GBM and intracerebral melanoma (174,175); (*3*) the HFR, Petten, The Netherlands and the University of Essen in Germany in 1997 using BSH (176); (*4*) the FiR1 at the Helsinki University Central Hospital (142) in 1999 to the present; (*5*) the Studsvik reactor facility in Sweden from 2001 to June 2005, carried out by the Swedish National Neuro-Oncology Group (141), and finally (*6*) the NRI reactor in Rez, Czech Republic by Tovarys using BSH (177). The number of patients treated in this study is small and the followup is still rather short.

Initially, clinical studies using epithermal neutron beams were primarily Phase I safety and dose-ranging trials and a BNCT dose to a specific volume or critical region of the normal brain was prescribed. In both the BNL and the Harvard/MIT clinical trials, the peak dose delivered to a $1\ cm^3$ volume was escalated in a systematic way. As the dose escalation trials have progressed, the treatments have changed from single-field irradiations or parallel opposed irradiations, to multiple noncoplanar irradiation fields, arranged in order to maximize the dose delivered to the tumor. A consequence of this approach has been a concomitant increase in the average doses delivered to normal brain. The clinical trials at BNL and Harvard/MIT using BPA (compound **1**) and an epithermal neutron beam in the United States have now been completed.

Analysis of the Brookhaven and MIT Clinical Results

The BNL and Harvard/MIT studies have provided the most detailed data relating to normal brain tolerance following BNCT. A residual tumor volume of $60\ cm^3$ or greater lead to a greater incidence of acute CNS toxicity. This primarily was related to increased intracranial pressure, resulting from tumor necrosis and the associated cerebral edema (152,173,174). The most frequently observed neurological side effect associated with the higher radiation doses, other than the residual tumor volume-related effects, was radiation related somnolence (178). This is a well-recognized effect following whole brain photon irradiation (179), especially in children with leukemia or lymphoma, who have received CNS irradiation. However, somnolence is not a very well-defined radiation related endpoint because it frequently is diagnosed after tumor recurrence has been excluded. Therefore, it is not particularly well suited as a surrogate marker for normal tissue tolerance. In the dose escalation studies carried out at BNL (152,173), the occurrence of somnolence in the absence of a measurable tumor dose response was clinically taken as the maximum tolerated normal brain dose. The volume-averaged whole brain

dose and the incidence of somnolence increased significantly as the BNL and Harvard/MIT trials progressed (175). The volume of tissue irradiated has been shown to be a determining factor in the development of side effects (180). Average whole brain doses greater than ~5.5 Gy(w) were associated with somnolence in the trial carried out at BNL, but not in all of the patients in the Harvard/MIT study (18,152,176). The BNL and Harvard/MIT trials were completed in 1999. Both produced median and 1-year survival times that were comparable to conventional external beam photon therapy (6). Although both were primarily Phase I trials to evaluate the safety of dose escalation as the primary endpoint for radiation related toxicity, the secondary endpoints were quality of life and time to progression and overall survival. The median survival times for 53 patients from the BNL trial and the 18 GBM patients from the Harvard/MIT trial were 13 months and 12 months, respectively. Following recurrence, most patients received some form of salvage therapy, which may have further prolonged overall survival. Time to progression, which would eliminate salvage therapy as a confounding factor, probably would be a better indicator of the efficacy of BNCT, although absolute survival time still is the "gold standard" for any clinical trial. The quality of life for most of the BNL patients was very good, especially considering that treatment was given in one or two consecutive daily fraction(s).

Clinical Trials Carried Out in Sweden and Finland

The clinical team at the Helsinki University Central Hospital and VTT (Technical Research Center of Finland) have reported on 18 patients using BPA as the capture agent (290 mg·kg^{-1} infused over 2 h) with two irradiation fields and whole brain average doses in the range of 3–6 Gy(w) (142). The estimated 1-year survival was 61%, which was very similar to the BNL data. This trial is continuing and the dose of BPA has been escalated to 450 mg·kg^{-1} and will be increased to 500 mg·kg^{-1}, infused over 2 h (H. Joensuu, personal communication). Since BNCT can deliver a significant dose to tumor with a relatively low average brain dose, this group also has initiated a clinical trial for patients who have recurrent GBM after having received full-dose photon therapy. In this protocol, at least 6 months must have elapsed from the end of photon therapy to the time of BNCT and the peak brain dose should be <8 Gy(w) and the whole brain average dose <6 Gy(w). As of August 2005, only a small number of patients have been treated, but this has been well tolerated.

Investigators in Sweden have carried out a BPA-based trial using an epithermal neutron beam at the Studsvik Medical AB reactor (141). This study differed significantly from all previous clinical trials in that the total amount of BPA administered was increased to 900 mg·kg^{-1}, infused i.v. over 6 h. This approach was based on the following preclinical data: (*1*) the *in vitro* observation that several hours were required to fully load cells with BPA (181); (*2*) long-term i.v. infusions of BPA in rats increased the absolute tumor boron concentrations in the 9L gliosarcoma model, although the T:Bl ratio remained constant (182,183), and (*3*) most importantly, long-term i.v. infusions of BPA appeared to improve the uptake of boron in infiltrating tumor cells at some distance from the main tumor mass in rats bearing intracerebral 9L gliosarcomas (184). The longer infusion time of BPA was well tolerated (185–187) by the 30 patients who were enrolled in this study. All patients were treated with two fields, and the average weighted whole brain dose was 3.2–6.1 Gy(w), which was lower than the higher end of the doses used in the Brookhaven trial, and the minimum dose to the tumor ranged from 15.4 to 54.3 Gy(w). At 10 months following BNCT 23 of 29 evaluable patients had died with a median time to progression following BNCT of 5.8 months and a median survival time of 14.2 months. These results are comparable but not better than those obtained with external beam radiation therapy. Furthermore, they emphasize the need to improve the delivery of BPA, as well as BSH. As part of a broader plan to restructure the company, a decision was made by Studsvik AB in June 2005 to terminate operation of both the R2-0 reactor, which was used for this clinical trial, and the R2 reactor.

CLINICAL STUDIES OF BNCT FOR OTHER TUMORS

Treatment of Melanoma

Other than patients with primary brain tumors, the second largest group that has been treated by BNCT were those with cutaneous melanomas. Mishima and co-workers previously had carried out extensive studies in experimental animals with either primary or transplantable melanomas using ^{10}B enriched BPA as the capture agent (188,189). The use of BPA was based on the premise that it would be selectively taken up by and accumulate in neoplastic cells that were actively synthesizing melanin (190). Although it was subsequently shown that a variety of malignant cells preferentially took up large amounts of BPA compared to normal cells (191), nevertheless, Mishima's studies clearly stimulated clinical interest in BPA as a boron delivery agent. Since BPA itself has low water solubility, it was formulated with HCl to make it more water soluble. The first patient, who was treated by Mishima in 1985, had an acral lentigenous melanoma of his right toe that had been amputated (192). However, 14 months later he developed a subcutaneous metastatic nodule on the left occiput, which was determined to be inoperable due to its location. The tumor was injected peritumorally at multiple points for a total dose of 200 mg of BPA. Several hours later, by which time BPA had cleared from normal skin, but still had been retained by the melanoma, the tumor was irradiated with a collimated beam of thermal neutrons. Based on the tumor boron concentrations and the neutron fluence, an estimated 45 RBE-Gy equivalent dose was delivered to the melanoma. Marked regression was noted after 2 months, and the tumor had completely disappeared by 9 months (188,189,192). This successful outcome provided further evidence for *proof-of-principle* of the usefulness of BNCT to treat a radioresistant tumor. Subsequently, at least an additional 18 patients with either primary or metastatic melanomas have been treated by Mishima and co-workers (193). The BPA either was injected peritumorally or administered orally as a slurry (194) until Yoshino et al.

improved its formulation and water solubility by complexing it with fructose, following which it was administered i.v. (195). This important advance ultimately led to the use of BPA in the clinical trials in patients with brain tumors that were described in the preceding section. In all of Mishima's patients, there was local control of the treated primary or metastatic melanoma nodule(s) and several patients were tumor free at 4 or more years following BNCT (193).

Several patients with either cutaneous or cerebral metastases of melanoma have been treated by Busse et al. using BPA fructose as the delivery agent (18,196). The most striking example of a favorable response was in a patient with an unresected cerebral metastasis in the occipital lobe. The tumor received a dose of 24 RBE-Gy and monthly MRI studies revealed complete regression over a 4 month interval (196). As evidenced radiographically, a second patient with a brain metastasis had a partial response. Several other patients with either cutaneous or metastatic melanoma to the brain have been treated at other institutions, including the first in Argentina (197), and the consensus appears to be that these tumors are more responsive to BNCT than GBMs. This is supported by experimental studies carried out by two of us (R.F.B. and J.A.C.) using a human melanoma xenograft model (198,199), which demonstrated enhanced survival times and cure rates superior to those obtained using the F98 rat glioma model (200). In summary, multicentric metastatic brain tumors, and more specifically melanoma, which cannot be treated either by surgical excision or stereotactic radiosurgery, may be candidates for treatment by BNCT.

Other Tumor Types Treated by BNCT

Two other types of cancer recently have been treated by BNCT. The first is recurrent tumors of the head and neck. Kato et al. reported on a series of six patients, three of whom had squamous cell carcinomas, two had sarcomas, and 1 had a parotid tumor (201). All of them had received standard therapy and had developed recurrent tumors for which there were no other treatment options. All of the patients received a combination of BSH (5 g) and BPA ($250\,mg{\cdot}kg^{-1}$ body weight), administered i.v. In all but one patient, BNCT was carried out at the Kyoto University Research Reactor using an epithermal neutron beam in one treatment that was given 12h following administration of BSH and 1 h after BPA. The patient with the parotid tumor, who received a second treatment one month following the first, had the best response with a 63% reduction in tumor volume at 1 month and a 94% reduction at 1 year following the second treatment without evidence of recurrence. The remaining five patients showed responses ranging from a 10–27% reduction in tumor volume with an improvement in clinical status. This study has extended the use of BNCT to a group of cancers that frequently are ineffectively treated by surgery, radio-, and chemotherapy. However, further clinical studies are needed to objectively determine the clinical usefulness of BNCT for head and neck cancers, and another study to assess this currently is in progress at Helsinki University Central Hospital.

The second type of tumor that recently has been treated by BNCT is adenocarcinoma of the colon that had metastasized to the liver (202). Although hepatectomy followed by allogeneic liver transplantation, has been carried out at a number of centers (203,204), Pinelli and Zonta et al. (202) in Pavia, Italy, have approached the problem of multicentric hepatic metastases using an innovative, but highly experimental procedure. Their patient had >14 metastatic nodules in the liver parenchyma, the size of which precluded surgical excision. Before hepatectomy was performed, the patient received a 2 h infusion of BPA fructose ($300\,mg{\cdot}kg^{-1}$ b.w.) via the colic vein. Samples of tumor and normal liver were taken for boron determinations and once it was shown that boron selectively had localized in the tumor nodules with small amounts in normal liver, the hepatectomy was completed (202). The liver then was transported to the Reactor Laboratory of the University of Pavia for neutron irradiation, following which it was reimplanted into the patient. More than 2 years later in October 2004, the patient had no clinical or radiographic evidence of recurrence and CEA levels were low (205). Although it is unlikely that this approach will have any significant clinical impact on the treatment of the very large number of patients who develop hepatic metastases from colon cancer, it nevertheless again provides *proof of principle* that BNCT can eradicate multicentric deposits of tumor in a solid organ. The Pavia group has plans to treat other patients with metastatic liver cancer and several other groups (206–208) are exploring the possibility of treating patients with primary, as well as metastatic tumors of the liver using this procedure.

CRITICAL ISSUES

There are a number of critical issues that must be addressed if BNCT is to become a useful modality for the treatment of cancer, and most specifically, brain tumors. *First* and foremost, there is a need for more selective and effective boron agents, which when used either alone or in combination, could deliver the requisite amounts ($\sim 20\,\mu g{\cdot}g^{-1}$) of boron to the tumor. Furthermore, their delivery must be optimized in order to improve both tumor uptake and cellular microdistribution, especially to different subpopulations of tumor cells (185). A number of studies have shown that there is considerable patient-to-patient as well-intratumoral variability in the uptake of both BSH (209,210) and BPA (184,211,212). At this point in time, the dose and delivery of these drugs have yet to be optimized, but based on experimental animal data (31,32,34,137,183), improvement in dosing and delivery could have a significant impact on increasing tumor uptake and microdistribution.

Second, since the radiation dosimetry for BNCT is based on the microdistribution of ^{10}B (209,213), which is indeterminable on a real-time basis, methods are needed to provide semiquantatative estimates of the boron content in the residual tumor. Imahori and co-workers (214–216) in Japan and Kabalka (217) in the United States have carried out imaging studies with ^{18}F-labeled BPA, and have used to establish the feasibility of carrying out BNCT. This

^{18}F-PET imaging also has been used as a prognostic indicator for patients with GBM who may or may not have received BNCT (214,215). In the former group, it has been used to establish the feasibility of carrying out BNCT based on the uptake and distribution of ^{18}F-BPA within the tumor and in the latter to monitor the response to therapy. The possibility of using MRI for either ^{10}B or ^{11}B has been under investigation (218), and this may prove to be useful for real-time localization of boron in residual tumor prior to BNCT. Magnetic resonance spectroscopy (MRS) and magnetic resonance spectroscopic imaging (MRSI) also may be useful for monitoring the response to therapy (219). Kojimoto and Miyatake et al. recently used MRS to analyze the target specificity of BPA and the effects of BNCT in a group of six patients using multivoxel proton MRS (220). There was a reduction in the choline/creatine ratio without a reduction of the *N*-acetylaspartate/creatine ratio at 14 days following BNCT, strongly suggesting that there was selective destruction of tumor cells and a sparing of normal neurons (220). Noninvasive procedures (e.g., MRSI) may be a powerful way to follow the clinical response to BCNT in addition to MRI. However, in the absence of real-time tumor boron uptake data, the dosimetry for BNCT is very problematic. This is evident from the discordance of estimated doses of radiation delivered to the tumor and the therapeutic response, which would have been greater than that which was seen if the tumor dose estimates were correct (152).

Third, there is a discrepancy between the theory behind BNCT, which is based on a very sophisticated concept of selective cellular and molecular targeting of high LET radiation, and the implementation of clinical protocols, which are based on very simple approaches to drug administration, dosimetry, and patient irradiation. This in part is due to the fact that BNCT has not been carried out in advanced medical settings with a highly multidisciplinary clinical team in attendance. At this time BNCT has been totally dependent on nuclear reactors as neutron sources. These are a medically unfriendly environment and are located at sites at varying distances from tertiary care medical facilities, which has made it difficult to attract patients, and the highly specialized medical team that ideally should be involved in clinical BNCT. Therefore, there is an urgent need for either very compact medical reactors or ABNS that could be easily sited at selected centers that treat large numbers of patients with brain tumors.

Fourth, there is a need for randomized clinical trials. This is especially important since almost all major advances in clinical cancer therapy have come from these, and up until this time no randomized trials of BNCT have been conducted. The pitfalls of nonrandomized clinical trials for the treatment of brain tumors have been well documented (221,222). It may be somewhat wishful thinking to believe that the clinical results with BNCT will be so clearcut that a clear determination of efficacy could be made without such trials. These will require a reasonably large number of patients in order to provide unequivocal evidence of efficacy with survival times significantly better than those obtainable with promising currently available therapy for both GBMs (223,224) and metastatic brain tumors (225). This leads to the issue of conducting such trials, which might best be accomplished through cooperative groups such as the Radiation Therapy Oncology Group (RTOG) in the United State or the European Organization for Research Treatment of Cancer (EORTC).

Finally, there are several promising leads that could be pursued. The upfront combination of BNCT with external beam radiation therapy or in combination with chemotherapy has not been explored, although recently published experimental data, suggest that there may be a significant gain if BNCT is combined with photon irradiation (34). The extension of animal studies, showing enhanced survival of brain tumor bearing rats following the use of BSH and BPA in combination, administered intraarterially with or without BBB-D, has not been evaluated clinically. This approach is promising, but it is unlikely that it could be carried out at a nuclear reactor.

As is evident from this article, BNCT represents an extraordinary joining together of nuclear technology, chemistry, biology, and medicine to treat cancer. Sadly, the lack of progress in developing more effective treatments for high grade gliomas has been part of the driving force that continues to propel research in this field. BNCT may be best suited as an adjunctive treatment, used in combination with other modalities, including surgery, chemotherapy, and external beam radiation therapy, which, when used together, may result in an improvement in patient survival. Clinical studies have demonstrated the safety of BNCT. The challenge facing clinicians and researchers is how to get beyond the current impasse. We have provided a road map to move forward, but its implementation still remains a daunting challenge

ACKNOWLEDGMENTS

We thank Mrs. Michelle Smith for secretarial assistance in the preparation of this manuscript. Text and the two figures, which appear in this article, have been published in *Clinical Cancer Research* with copyright release from the American Association of Cancer Research, Inc.

Experimental studies described in this article have been supported by the National Institutes of Health grants 1R01 CA098945 (to R.F.B.) and 1R01 CA098902 to (M.G.H.V.) and Department of Energy Grants DE-FG02-93ER61612 (to T.E.B.) and DE-FG02-01ER63194 (to J.A.C.) and the Royal G. and Mae H. Westaway Family Memorial Fund at the Massachusetts Institute of Technology (to J.A.C.).

BIBLIOGRAPHY

Cited References

1. Berger MS. Malignant astrocytomas: surgical aspects. Seminars Oncol 1994;21:172–185.
2. Gutin PH, Posner JB. Neuro-Oncology: diagnosis and management of cerebral gliomas—past, present, and future. Neurosurgery 2000;47:1–8.
3. Parney IF, Chang SM. Current chemotherapy for glioblastoma. In: Market J, DeVita VT, Rosenberg SA, Hellman S, editors. Glioblastoma Multiforme, 1st ed., Sudbury (MA): Jones and Bartlett Publishers; 2005. p 161–177.

4. Paul DB, Kruse CA. Immunologic approaches to therapy for brain tumors. Curr Neurol Neurosci Rep 2001;1:238–244.
5. Rainov NG. Gene therapy for human malignant brain tumors. In: Market J, DeVita VT, Rosenberg SA, Hellman S, editors. Glioblastoma Multiforme. 1st ed. Sudbury (MA): Jones and Bartlett Publishers; 2005. p 249–265.
6. Curran WJ, et al. Recursive partitioning analysis of prognostic factors in three radiation oncology group malignant glioma trials. J Nat Cancer Inst 1993;85:704–710.
7. Lacroix M, et al. A multivariate analysis of 416 patients with glioblastoma multiforme: prognosis, extent of resection, and survival. J Neurosurg 2001;95:190–198.
8. Hentschel SJ, Lang FF. Current surgical management of glioblastoma. In: Market J, DeVita VT, Rosenberg SA, Hellman S, editors. Glioblastoma Multiforme, 1st ed., Sudbury (MA): Jones and Bartlett Publishers; 2005. p 108–130.
9. Laws ER, Shaffrey ME. The inherent invasiveness of cerebral gliomas: implications for clinical management. Int J Devel Neurosc 1999;17:413–420.
10. Halperin EC, Burger PC, Bullard DE. The fallacy of the localized supratentorial malignant glioma. Int J Radiat Oncol Biol Phys 1988;15:505–509.
11. Kaczarek E, et al. Dissecting glioma invasion: interrelation of adhesion, migration and intercellular contacts determine the invasive phenotype. Int J Devel Neurosc 1999;17:625–641.
12. Huang S, Prabhu S, Sawaya R. Molecular and biological determinants of invasiveness and angiogenesis in central nervous system tumors. In: Zhang W, Fuller GN, editors. Genomic and Molecular Neuro-Oncology. Sudbury (MA): Jones and Bartlette Publishers; 2004. pp 97–118.
13. Parney IF, Hao C, Petruk K. Glioma immunology and immunotherapy. Neurosurgery 2000;46:778–792.
14. Ware ML, Berger MS, Binder DK. Molecular biology of glioma tumorigenesis. Histol Histopathol 2003;18:207–216.
15. Barth RF. A critical assessment of boron neutron capture therapy: An overview. J Neuro-Oncol 2003;62:1–5.
16. Mishima Y. Selective thermal neutron capture therapy of cancer cells using their specific metabolic activities - melanoma as prototype. In: Mishima Y, editor. Cancer Neutron Capture Therapy. New York: Plenum Press; 1996. p 1–26.
17. Busse PM, et al. A critical examination of the results from the Harvard-MIT NCT program phase I clinical trial of neutron capture therapy for intracranial disease. J Neuro-Oncol 2003;62:111–121.
18. Coderre JA, et al. Boron neutron capture therapy: cellular targeting of high linear energy transfer radiation. Technol Cancer Res Treatment 2003;2:1–21.
19. Sauerwein W, Moss R, Wittig A, editors. Research and development in neutron capture therapy. Bologna, Italy: Monduzzi Editore S.p.A., International Proceedings Division; 2002.
20. Coderre JA, Rivard MJ, Patel H, Zamenhof RG. Proceedings of the 11th World Congress on Neutron Capture Therapy. Appl Rad Isotopes 2004; 61s.
21. Coderre JA, Morris GM. The radiation biology of boron neutron capture therapy. Radiat Res 1999;151:1–18.
22. Morris GM, et al. Response of the central nervous system to boron neutron capture irradiation: Evaluation using rat spinal cord model. Radiother Oncol 1994;32:249–255.
23. Morris GM, et al. Response of rat skin to boron neutron capture therapy with *p*-boronophenylalanine or borocaptate sodium. Radiother Oncol 1994;32:144–153.
24. Gupta N, Gahbauer RA, Blue TE, Albertson B. Common challenges and problems in clinical trials of boron neutron capture therapy of brain tumors. J Neuro-Oncol 2003;62:197–210.
25. Nigg DW. Computational dosimetry and treatment planning considerations for neutron capture therapy. J Neuro-Oncol 2003;62:75–86.
26. Coderre JA, et al. Boron neutron capture therapy for glioblastoma multiforme using p-boronophenylalanine and epithermal neutrons: Trial design and early clinical results. J Neuro-Oncol 1997;33:141–152.
27. Elowitz EH, et al. Biodistribution of p-boronophenylalanine in patients with glioblastoma multiforme for use in boron neutron capture therapy. Neurosurgery 1998;42:463–469.
28. Blue TE, Yanch JC. Accelerator-based epithermal neutron sources for boron neutron capture therapy of brain tumors. J Neuro-Oncol 2003;62:19–31.
29. Fukuda H, et al. Boron neutron capture therapy of malignant melanoma using ^{10}B-paraboronophenylalanine with special reference to evaluation of radiation dose and damage to the skin. Radiat Res 1994;138:435–442.
30. Coderre JA, et al. Derivations of relative biological effectiveness for the high-LET radiations produced during boron neutron capture irradiations of the 9L rat gliosarcoma *in vitro* and *in vivo*. Int J Radiat Oncol Biol Phys 1993;27: 1121–1129.
31. Barth RF, et al. Boron neutron capture therapy of brain tumors: enhanced survival following intracarotid injection of either sodium borocaptate or boronophenylalanine with or without blood-brain barrier disruption. Cancer Res 1997;57: 1129–1136.
32. Barth RF, et al. Boron neutron capture therapy of brain tumors: enhanced survival and cure following blood-brain barrier disruption and intracarotid injection of sodium borocaptate and boronophenylalanine. Int J Radiat Oncol Biol Phys 2000;47:209–218.
33. Miyatake S, et al. Modified boron neutron capture therapy (BNCT) for mailignant gliomas using epithermal neutrons and two boron compounds with different accumulation mechanisms-Effectiveness of BNCT on radiographic images. J Neurosurg Dec. 2005 (In Press).
34. Barth RF, et al. Combination of boron neutron capture therapy and external beam X-irradiation for the treatment of brain tumors. Int J Radiat Oncol Biol Phys 2004;58: 267–277.
35. Farr LE, et al. Neutron capture therapy with boron in the treatment of glioblastoma multiforme. Am J Roenthemol 1954;71:279–291.
36. Godwin JT, Farr LE, Sweet WH, Robertson JS. Pathological study of eight patients with glioblastoma multiforme treated by neutron-capture therapy using boron 10. Cancer 1955;8:601–615.
37. Snyder HR, Reedy AJ, Lennarz WJ. Synthesis of aromatic boronic acids, aldehydo boronic acids and a boronic acid analog of tyrosine. J Am Chem Soc 1958;80:835–838.
38. Soloway AH, Hatanaka H, Davis MA. Penetration of brain and brain tumor. VII. Tumor-binding sulfhydryl boron compounds. J Med Chem 1967;10:714.
39. Hawthorne MF. The role of chemistry in the development of boron neutron capture therapy of cancer. Angew Chem Int Ed Engl 1993;32:950–984.
40. Morin C. The chemistry of boron analogues of biomolecules. Tetrahedron 1994;50:12521–12569.
41. Soloway AH, et al. The chemistry of neutron capture therapy. Chem Rev 1998;98:1515–1562.
42. Hawthorne MF, Lee MW. A critical assessment of boron target compounds for boron neutron capture therapy. J Neuro-Oncol 2003;62:33–45.
43. Olsson P, et al. Uptake of a boronated epidermal growth factor-dextran conjugate in CHO xenografts with and without human EGF-receptor expression. Anticancer Drug Des 1998;13:279–289.
44. Gabel D, Foster S, Fairchild RG. The Monte Carlo simulation of the biological effect of the ^{10}B(n,α)^{7}L reaction in cells and

tissue and its implication for boron neutron capture therapy. Radiat Res 1987;111:14–25.
45. Srivastava RR, Singhaus RR, Kabalka GW. 4-Dihydroxyborylphenyl analogues of 1-aminocyclobutanecarboxylic acids: potential boron neutron capture therapy agents. J Org Chem 1999;64:8495–8450.
46. Das BC, et al. Synthesis of a water soluble carborane containing amino acid as a potential therapeutic agent. Syn Letteve 2001;9:1419–1420.
47. Kabalka GW, Yao M-L. Synthesis of a novel boronated 1-amino-cyclobutanecarboxylic acid as a potential boron neutron capture therapy agent. App Organomet Chem 2003;17: 398–402.
48. Diaz S, Gonzalez A, De Riancho SG, Rodriguez A. Boron complexes of *S*-trityl-L-cysteine and *S*-tritylglutathione. J Organomet Chem 2000;610:25–30.
49. Lindström P, Naeslund C, Sjöberg S. Enantioselective synthesis and absolute configurations of the enantiomers of o-carboranylalanine. Tetrahedron Lett 2000;41:751–754.
50. Masunaga S-I, et al. Potential of α-amino alcohol p-boronophenylalaninol as a boron carrier in boron neutron capture therapy, regarding its enantiomers. J Cancer Res Clin Oncol 2003;129:21–28.
51. Diaz A, Stelzer K, Laramore G, Wiersema R. Pharmacology studies of $Na_2\ ^{10}B_{10}H_{10}$ (GB-10) in human tumor patients. In: Sauerwein W, Moss R, Wittig A, editors. Research and Development in Neutron Capture Therapy. Bologna : Monduzzi Editore, International Proceedings Division; 2002. p 993–999.
52. Hawthorne MF, Feakes DA, Shelly K. Recent results with liposomes as boron delivery vehicles from boron neutron capture therapy. In: Mishima Y, editor Cancer Neutron Capture Therapy. New York: Plenum Press; 1996. p 27–36.
53. Feakes DA, Waller RC, Hathaway DK, Morton VS. Synthesis and in vivo murine evaluation of $Na_4[1\text{-}(1'\text{-}B_{10}H_9)\text{-}6\text{-}SHB_{10}H_8]$ as a potential agent for boron neutron capture therapy. Proc Natl Acad Sci USA 1999;96:6406–6410.
54. Shukla S, et al. Evaluation of folate receptor targeted boronated starburst dendrimer as a potential targeting agent for boron neutron capture therapy. Bioconjugate Chem 2003;14:158–167.
55. Sudimack J, et al. Intracellular delivery of lipophilic boron compound using folate receptor-targeted liposomes. Pharm Res 2002;19:1502–1508.
56. Takagaki M, et al. Boronated dipeptide borotrimethylglycylphenylalanine as a potential boron carrier in boron neutron capture therapy for malignant brain tumors. Radiat Res 2001;156:118–122.
57. Wakamiya T, et al. Synthesis of 4-boronophenylalanine-containing peptides for boron neutron capture therapy of cancer cells. Peptide Sci 1999;36:209–212.
58. Lesnikowski ZJ, Schinazi RF. Boron containing oligonucleotides. Nucleosides Nucleotides 1998;17:635–647.
59. Soloway AH, et al. Identification, development, synthesis and evaluation of boron-containing nucleosides for neutron capture therapy. J Organomet Chem 1999;581:150–155.
60. Lesnikowski ZJ, Shi J, Schinazi RF. Nucleic acids and nucleosides containing carboranes. J Organo-met Chem 1999;581: 156–169.
61. Lunato AJ, et al. Synthesis of 5-(carboranylalkylmercapto)-2′-deoxyuridines and 3-(carboranylalkyl)thymidines and their evaluation as substrates for human thymidine kinases 1 and 2. J Med Chem 1999;42:3378–3389.
62. Al-Madhoun AS, et al. Synthesis of a small library of 3-(carboranylalkyl)thymidines and their biological evaluation as substrates for human thymidine kinases 1 and 2. J Med Chem 2002;45:4018–4028.
63. Schinazi RF, et al. Treatment of Isografted 9L rat brain tumors with *b*-5-o-carboranyl-2'-deoxyuridine neutron capture therapy. Clin Cancer Res 2000;6:725–730.
64. Al-Madhoun AS, et al. Evaluation of human thymidine kinase 1 substrates as new candidates for boron neutron capture therapy. Cancer Res 2004;64:6280–6286.
65. Barth RF, et al. Boron containing nucleosides as potential delivery agents for neutron capture therapy of brain tumors. Cancer Res 2004;64:6287–6295.
66. Sjöberg S, et al. Chemistry and biology of some low molecular weight boron compounds for boron neutron capture therapy. J Neuro-Oncol 1997;33:41–52.
67. Tietze LF, et al. Novel carboranes with a DNA binding unit for the treatment of cancer by boron neutron capture therapy. ChemBio-Chem 2002;3:219–225.
68. Bateman SA, Kelly DP, Martin RF, White JM. DNA binding compounds. VII. Synthesis, characterization and DNA binding capacity of 1,2-dicarba-*closo*-dodecaborane bibenzimidazoles related to the DNA minor groove binder Hoechst 33258. Aust J Chem 1999;52:291–301.
69. Woodhouse SL, Rendina LM. Synthesis and DNA-binding properties of dinuclear platinum(II)-amine complexes of 1,7-dicarba-*closo*-dodecaborane(12). Chem Commun 2001; 2464–2465.
70. Cai J, et al. Boron-containing polyamines as DNA-targeting agents for neutron capture therapy of brain tumors: synthesis and biological evaluation. J Med Chem 1997;40:3887–3896.
71. Zhuo J-C, et al. Synthesis and biological evaluation of boron-containing polyamines as potential agents for neutron capture therapy of brain tumors. J Med Chem 1999;42: 1281–1292.
72. Martin B, et al. *N*-Benzylpolyamines as vectors of boron and fluorine for cancer therapy and imaging: synthesis and biological evaluation. J Med Chem 2001;44:3653–3664.
73. El-Zaria ME, Doerfler U, Gabel D. Synthesis of [(amino-alkylamine)-*N*-amino-alkyl] azanonaborane(11) derivatives for boron neutron capture therapy. J Med Chem 2002;45: 5817–5819.
74. Nakanishi A, et al. Toward a cancer therapy with boron-rich oligomeric phosphate diesters that target the cell nucleus. Proc Natl Acad Sci USA 1999;96:238–241.
75. Maderna A, et al. Synthesis of a porphyrin-labelled carboranyl phosphate diester: a potential new drug for boron neutron capture therapy of Cancer. Chem Commun 2002; 1784–1785.
76. Vicente MGH. Porphyrin-based sensitizers in the detection and treatment of cancer: recent progress. Curr Med Chem Anti-Cancer Agents 2001;1:175–194.
77. Bregadze VI, Sivaev IB, Gabel D, Wöhrle D. Polyhedral boron derivatives of porphyrins and phthalocyanines. J Porphyrins Phthalocyanines 2001;5:767–781.
78. Evstigneeva RP, et al. Carboranylporphyrins for boron neutron capture therapy of cancer. Curr Med Chem: Anti-Cancer Agents 2003;3:383–392.
79. Vicente MGH, et al. Syntheses, toxicity and biodistribution of two 5,15-di[3,5-(*nido*-carboranyl-methyl)phenyl] porphyrin in EMT-6 tumor bearing mice. Bioorg Med Chem 2003;11: 3101–3108.
80. Miura M, et al. Evaluation of carborane-containing prophyrins as tumour agents for boron neutron capture therapy. Br J Radiol 1998;71:773–781.
81. Miura M, et al. Biodistribution of copper carboranyltetraphenylporphyrins in rodents bearing an isogeneic or human neoplasm. J Neuro-Oncol 2001;52:111–117.
82. Miura M, et al. Boron neutron capture therapy of a murine mammary carcinoma using a lipophilic carboranyltetraphenylporphyrin. Radiat Res 2001;155:603–610.

83. Gottumukkala V, Luguya R, Fronczek FR, Vicente MGH. Synthesis and cellular studies of an octa-anionic 5,10,15,20-tetra[3,5-(*nido*-carboranylmethyl)phenyl] porphyrin (H_2OCP) for application in BNCT. Bioorg Med Chem 2005;13: 1633–1640.
84. Hao E, Vicente MGH. Expeditious synthesis of porphyrin-cobaltacarborane vonjugates. Chem Commun 2005; 1306–1308.
85. Ongayi O, Gottumukkala V, Fronczek FR, Vicente MGH. Synthesis and characterization of a carboranyl-tetrabenzoporphyrin. Bioorg Med Chem Lett 2005;15:1665–1668.
86. Fabris C, Jori G, Giuntini F, Roncucci G. Photosensitizing properties of a boronated phthalocyanine: studies at the molecular and cellular level. J Photochem Photobiol B: Biol 2001;64:1–7.
87. Giuntini F, et al. Synthesis of tetrasubstitited Zn(II)-phthalocyanines carrying four carboranyl-units as potential BNCT and PDT agents. Tetrahedron Lett 2005;46:2979–2982.
88. Luguya R, Fronckzek FR, Smith KM, Vicente MGH. Synthesis of novel carboranylchlorins with dual application in boron neutron capture therapy (BNCT) and photodynamic therapy (PDT). Appl Rad Isotopes 2004;61:1117–1123.
89. Rosenthal MA, Kavar B, Uren S, Kaye AH. Promising survival in patients with high-grade gliomas following therapy with a novel boronated porphyrin. J Clin Neurosc 2003; 10:425–427.
90. Rosenthal MA, et al. Phase I and pharmacokinetic study of photodynamic therapy for high-grade gliomas using a novel boronated porphyrin. J Clin Oncol 2001;19:519–524.
91. Hill JS, et al. Selective tumor kill of cerebral glioma by photodynamic therapy using a boronated phorhyrin photosensitizer. Proc Natl Acad Sci USA 1995;92:12126–12130.
92. Kawabata S, et al. Evaluation of the carboranyl porphyrin H_2TCP as a delivery agent for boron neutron capture therapy (BNCT). Khamlichi A, editor. 13th World Congress of Neurological Surgery, Marakesh, Morocco. June 19–24, 2005. p 975–979.
93. Ozawa T, et al. *In vivo* evaluation of the boronated porphyrins TABP-1 in U-87 MG intracerebral human glioblastoma xenografts. Mol Pharmaceut 2004;5:368–374.
94. Lauceri R, Purrello R, Shetty SJ, Vicente MGH. Interactions of anionic carboranylated porphyrins with DNA. J Am Chem Soc 2001;123:5835–5836.
95. Vicente MGH, et al. Synthesis, dark toxicity and induction of in vitro DNA photodamage by a tetra(4-nido-carboranylphenyl)porphyrin. J Photochem Photobiol B: Biol 2002;68:123–132.
96. Ghaneolhosseini H, Tjarks W, Sjoberg S. Synthesis of novel boronated acridines and spermidines as possible agents for BNCT. Tetrahedron 1998;54:3877–3884.
97. Gedda L, et al. Cytotoxicity and subcellular localization of boronated phenanthridinium analogs. Anti-Cancer Drug Design 1997;12:671–685.
98. Gedda L, et al. The influence of lipophilicity on binding of boronated DNA-intercalating compounds in human glioma spheroids. Anti-Cancer Drug Design 2000;15:277–286.
99. Giovenzana GB, et al. Synthesis of carboranyl derivatives of alkynyl glycosides as potential BNCT agents. Tetrahedron 1999;55:14123–14136.
100. Tietze LF, et al. Ortho-carboranyl glycosides for the treatment of cancer by boron neutron capture therapy. Bioorg Med Chem 2001;9:1747–1752.
101. Orlova AV, et al. Conjugates of polyhedral boron compounds with carbohydrates. 1. New approach to the design of selective agents for boron neutron capture therapy of cancer. Russ Chem Bull 2003;52:2766–2768.
102. Tietze LF, Bothe U. Ortho-carboranyl glycosides of glucose, mannose, maltose and lactose for cancer treatment by boron neutron-capture therapy. Chem Eur J 1998;4:1179–1183.
103. Raddatz S, et al. Synthesis of new boron-rich building blocks for boron neutron capture therapy or energy-filtering transmission electron microscopy. ChemBioChem 2004;5:474–482.
104. Tietze LF, et al. Novel carboranyl C-glycosides for the treatment of cancer by boron neutron capture therapy. Chem Eur J 2003;9:1296–1302.
105. Basak P, Lowary TL. Synthesis of conjugates of L-fucose and *ortho*-carborane as potential agents for boron neutron capture therapy. Can J Chem 2002;80:943–948.
106. Endo Y, et al. Structure–activity study of estrogenic agonists bearing dicarba-*closo*-dodecaborane. Effect of geometry and separation distance of hydroxyl groups at the ends of molecules. Bioorg Med Chem Lett 1999;9:3313–3318.
107. Lee J-D, et al. A convenient synthesis of the novel carboranyl-substituted tetrahydroisoquinolines: application to the biologically active agent for BNCT. Tetrahedron Lett 2002;43:5483–5486.
108. Valliant JF, Schaffer P, Stephenson KA, Britten JF. Synthesis of Boroxifen, a *nido*-carborane analogue of tamoxifen. J Org Chem 2002;67:383–387.
109. Feakes DA, Spinler JK, Harris FR. Synthesis of boron-containing cholesterol derivatives for incorporation into unilamellar liposomes and evaluation as potential agents for BNCT. Tetrahedron 1999;55:11177–11186.
110. Endo Y, et al. Potent estrogen agonists based on carborane as a hydrophobic skeletal structure: a new medicinal application of boron clusters. Chem Biol 2001;8:341–355.
111. Tjarks W, et al. *In vivo* evaluation of phosphorous-containing derivatives of dodecahydro-*closo*-dodecaborate for boron neutron capture therapy of gliomas and sarcomas. Anticancer Res 2001;21:841–846.
112. Adams DM, Ji W, Barth RF, Tjarks W. Comparative *in vitro* evaluation of dequalinium B, a new boron carrier for neutron capture therapy (NCT). Anticancer Res 2000;20: 3395–3402.
113. Zakharkin LI, et al. Synthesis of bis(dialkylaminomethyl)-*o*- and *m*-carboranes and study of these compounds as potential preparations for boron neutron capture therapy. Pharm Chem J 2000;34:301–304.
114. Barth RF, et al. Boronated starburst dendrimer-monoclonal antibody immunoconjugates: evaluation as a potential delivery system for neutron capture therapy. Bioconjug Chem 1994;5:58–66.
115. Liu L, et al. Critical evaluation of bispecific antibodies as targeting agents for boron neutron capture therapy of brain tumors. Anticancer Res 1996;16:2581–2588.
116. Liu L, et al. Bispecific antibodies as targeting agents for boron neutron capture therapy of brain tumors. J Hematother 1995;4:477–483.
117. Novick S, et al. Linkage of boronated polylysine to glycoside moieties of polyclonal antibody; Boronated antibodies as potential delivery agents for neutron capture therapy. Nuclear Med Biol 2002;29:93–101.
118. Wu G, et al. Site-specific conjugation of boron containing dendrimers to anti-EGF receptor monoclonal antibody cetuximab (IMC-C225) and its evaluation as a potential delivery agent for neutron capture therapy. Bioconjugate Chem 2004;15:185–194.
119. Fallot T, et al. A Phase I study of an anti-epidermal growth factor receptor monoclonal antibody for the treatment of malignant gliomas. Neurosurgery 1996;39:478–483.

120. Carlsson J, et al. Strategy for boron neutron capture therapy against tumor cells with over-expression of the epidermal growth factor receptor. Int J Radiat Oncol Biol Phys 1994;30:105–115.
121. Capala J, et al. Boronated epidermal growth factor as a potential targeting agent for boron neutron capture therapy of brain tumors. Bioconjugate Chem 1996;7:7–15.
122. Sauter G, et al. Patterns of epidermal growth factor receptor amplification in malignant gliomas. Am J Pathol 1996;148: 1047–1053.
123. Schwechheimer K, Huang S, Cavenee WK. EGFR gene amplification-rearrangement in human glioblastoma. Int J Cancer 1995;62:145–148.
124. Backer MV, Backer JM. Targeting endothelial cells overexpressing VEGFR-2: selective toxicity of Shiga-like toxin-VEGF fusion proteins. Bioconjugate Chem 2001;12: 1066–1073.
125. Backer MV, et al. Vascular endothelial growth factor selectively targets boronated dendrimers to tumor vasculature. Mol Cancer Therapeut 2005;4:1423–1429.
126. Feakes DA, Shelly K, Hawthorne M. Selective boron delivery to murine tumors by lipophilic species incorporated in the membranes of unilamellar liposomes. Proc Natl Acad Sci USA 1995;92:1367–1370.
127. Carlsson J, et al. Ligand liposomes and boron neutron capture therapy. J Neuro-Oncol 2003;62:47–59.
128. Pardridge WM. Drug delivery to the brain. J Cerebral Blood Flow Metabol 1997;17:713–731.
129. Mendelsohn, J. Targeting the epidermal growth factor receptor for cancer therapy. J Clin Oncol 2002;20:1s–13s.
130. Nygren P, Sorbye H, Osterland P, Pfeiffer P. Targeted drugs in metastatic colorectal cancer with emphasis on guidelines for the use of bevacizumab and cetuximab. An Acta Oncologica expert report. Acta Oncolog 2005;44:203–218.
131. Wikstrand CJ, Cokgor I, Sampson JH, Bigner DD. Monoclonal antibody therapy of human gliomas: current status and future approaches. Cancer Metastasis Rev 1999;18: 451–464.
132. Barth RF, et al. Molecular targeting of the epidermal growth factor receptor for neutron capture therapy of gliomas. Cancer Res 2002;62:3159–3166.
133. Yang W, et al. Convection enhanced delivery of boronated epidermal growth factor for molecular targeting of EGFR positive gliomas. Cancer Res 2002;62:6552–6558.
134. Barth RF, et al. Neutron capture therapy of epidermal growth factor positive gliomas using boronated cetuximab (IMC-C225) as a delivery agent. App Radiat Isotopes 2004;61: 899–903.
135. Yang W, et al. Boronated epidermal growth factor as a delivery agent for neutron capture therapy of EGFR positive gliomas. App Rad Isotopes 2004;61:981–985.
136. Barth RF, et al. Enhanced delivery of boronophenylalanine for neutron capture therapy of brain tumors using the bradykinin analogue, Cereport™(RMP7). Neurosurgery 1999; 44:350–359.
137. Barth RF, et al. Neutron capture therapy of intracerebral melanoma: Enhanced survival and cure following blood–brain barrier opening to improve delivery of boronophenylalanine. Int J Radiat Oncol Biol Phys 2002;52:858–868.
138. Yang W, et al. Development of a syngeneic rat brain tumor model expressing EGFRvIII and its use for molecular targeting studies with monoclonal antibody L8A4. Clin Cancer Res 2005;11:341–350.
139. Harling O, Riley K. Fission reactor neutron sources for neutron capture therapy—a critical review. J Neuro-Oncol 2003;2:7–17.
140. Harling O, et al. The fission converter-based epithermal neutron irradiation facility at the Massachusetts Institute of Technology Reactor. Nuclear Sci Eng 2002;140:223–240.
141. Capala J, et al. Boron neutron capture therapy for glioblastoma multiforme: clinical studies in Sweden. J Neuro-Oncol 2003;62:135–144.
142. Joensuu H, et al. Boron neutron capture therapy of brain tumors: clinical trials at the Finnish Facility using boronophenylalanine. J Neuro-Oncol 2003;62:123–134.
143. Moss RL, et al. Design, construction and installation of an epithermal neutron beam for BNCT at the High Flux Reactor Petten. In: Allen BJ, et al., editors. Progress in Neutron Capture Therapy for Cancer, New York: Plenum Press; 1992. p 63–66.
144. Marek M, Viererbl M, Burian J, Jansky B. Determination of the geometric and spectral characteristics of BNCT beam (neutron and gamma-ray). In: Hawthorne MF, Shelly K, Wiersema RJ, editors., Neutron Capture Therapy, Vol. I, New York: Kluwer Academic/Plenum Publishers; 2001. p 381–389.
145. Kobayashi T, et al. The remodeling and basic characteristics of the heavy water neutron irradiation facility of the Kyoto University Research Reactor, Mainly for Neutron Capture Therapy. Nucl Technol 2000;131:354–378.
146. Yamamoto K, et al. Characteristics of neutron beams for BNCT. Proceedings of the 9th Symposium on Neutron Capture Therapy, Osaka, Japan, October 2–6, 2000. p 243–244.
147. Blaumann HR, Larrieu OC, Longhino JM, Albornoz AF. NCT facility development and beam characterisation at the RA-6 Reactor. In: Hawthorne MF, Shelly K, Wiersema RJ, editors. Frontiers in Neutron Capture Therapy. Vol. I, New York: Kluwer Academic/Plenum Publishers; 2001. p 313–317.
148. Agosteo S, et al. Design of neutron beams for boron neutron capture therapy in a fast reactor. IAEA Technical Committee Meeting about the Current Issues Relating to Neutron Capture Therapy, June 14–18, 1999, Vienna, Austria.
149. Fairchild RG, et al. Installation and testing of an optimized epithermal neutron beam at the Brookhaven Medical Research Reactor (BMRR). Proceedings of the Workshop on Neutron Beam Design, Development and Performance for Neutron Capture Therapy. MIT, Cambridge (MA), March 29–31, 1989.
150. Hatanaka H. Boron neutron capture therapy for brain tumors. In: Karin ABMF, Laws E, editor. Glioma: Berlin: Springer-Verlag; 1991. p 233–249.
151. Hatanaka H, Nakagawa Y. Clinical results of long-surviving brain tumor patients who underwent boron neutron capture therapy. Int J Radiat Oncol Biol Phys 1994;28:1061–1066.
152. Diaz AZ. Assessment of the results from the phase I/II boron neutron capture therapy trials at the Brookhaven National Laboratory from a clinician's point of view. J Neuro-Oncol 2003;62:101–109.
153. Riley K, Binns P, Harling O. Performance characteristics of the MIT fission converter based epithermal neutron beam. Phys Med Biol 2003;48:943–958.
154. Nigg D, et al. Initial neutronic performance assessment of an epithermal neutron beam for neutron capture therapy research at Washington State University. Research and Development in Neutron Capture Therapy. Proceedings of the 10th International Congress on Neutron Capture Therapy, 2002. p 135–139.
155. Beynon T, et al. Status of the Birmingham accelerator-based BNCT facility. Proceedings of the 10th International Congress on Neutron Capture Therapy, 2002. p 225–228.

156. Burlon A, et al. Optimization of a neutron production target and beam shaping assembly based on the $^{7}Li(p,n)^{7}Be$ reaction. Proceedings of the 10th International Congress on Neutron Capture Therapy, 2002. p 229–234.
157. Kononov O, et al. Investigations of using near-threshold $^{7}Li(p,n)^{7}Be$ reaction for NCT based on in-phantom dose distribution. Proceedings of the 10th International Congress on Neutron Capture Therapy, 2002. p 241–246.
158. Blackburn B, Yanch J, Klinkowstein R. Development of a high-power water cooled beryllium target for use in accelerator-based boron neutron capture therapy. Med Phys 1998;10:1967–1974.
159. Hawk A, Blue T, Woollard J, Gupta N. Effects of target thickness on neutron field quality for an ABNS. Research and Development in Neutron Capture Therapy Proceedings of the 10th International Congress on Neutron Capture Therapy, 2002. p 253–257.
160. Sakurai Y, Kobayashi T, Ono K. Study on accelerator-based neutron irradiation field aiming for wider application in BNCT - spectrum shift and regional filtering. Proceedings of the 10th International Congress on Neutron Capture Therapy, 2002. p 259–263.
161. Giusti V, Esposito J. Neutronic feasibility study of an accelerator-based thermal neutron irradiation cavity. Proceedings of the 10th International Congress on Neutron Capture Therapy, 2002. p 305–308.
162. Starling WJ. RFI Linac for accelerator-based neutrons. Abstracts of the 11th World Congress on Neutron Capture Therapy. Boston, October 11–15, 2004. p 45.
163. Locher GL. Biological effects and therapeutic possibilities of neutrons. Am J Roetgenol Radium Ther 1936;36:1–13.
164. Asbury AK, Ojeann, Nielson SL, Sweet WH. Neuropathologic study of fourteen cases of malignant brain tumor treated by boron-10 slow neutron capture therapy. J Neuropathol Exp Neurol 1972;31:278–303.
165. Sweet WH. Practical problems in the past in the use of boron-slow neutron capture therapy in the treatment of glioblastoma multiforme. Proceedings First International Symposium Neutron Capture Therapy, Brookhaven National Lab Reports 51730. October 12–14, 1983. p 376–378.
166. Slatkin DN. A history of boron neutron capture therapy of brain tumours. Postulation of a brain radiation dose tolerance limit. Brain 1991;114:1609–1629.
167. Nakagawa Y, Hatanaka H. Boron neutron capture therapy: Clinical brain tumor studies. J Neuro-Oncol 1997;33:105–115.
168. Nakagawa Y, et al. Clinical review of the Japanese experience with boron neutron capture therapy and a proposed strategy using epithermal neutron beams. J Neuro-Oncol 2003;62:87–99.
169. Laramore GE, et al. Boron neutron capture therapy: a mechanism for achieving a concomitant tumor boost in fast neutron radiotherapy. Int J Radiat Oncol Biol Phys 1994;28:1135–1142.
170. Kageji T, et al. Pharmacokinetics and boron uptake of BSH ($Na_2B_{12}H_{11}SH$) in patients with intracranial tumors. J Neuro-Oncol 1997;33:117–130.
171. Bergland R, et al. A Phase 1 trial of intravenous boronophenylalanine-fructose complex in patients with glioblastoma multiforme. Cancer Neutron Capture Therapy. Mishima Y, editor. New York: Plenum Press; 1996. p 739–746.
172. Coderre JA, et al. Biodistribution of boronophenylalanine in patients with glioblastoma multiforme: Boron concentration correlates with tumor cellularity. Radiat Res 1998; 149:163–170.
173. Chanana AD, et al. Boron neutron capture therapy for glioblastoma multiforme: interim results from the phase I/II dose-escalation studies. Neurosurgery 1999;44:1182–1193.
174. Busse PM, et al. A critical examination of the results from the Harvard-MIT NCT program phase I clinical trial of neutron capture therapy for intracranial disease. J Neuro-Oncol 2003;111–121.
175. Palmer MR, et al. Treatment planning and dosimetry for the Harvard-MIT phase I clinical trial of cranial neutron capture therapy. Int J Radiat Oncol Biol Phys 2002;53:1361–1379.
176. Wittig A, et al. Current clinical results of the EORTC-study 11961, in: Research and Development in Neutron Capture Therapy. Sauerwein W, Moss R, Wittig A, editors. Bologna: Monduzzi Editore; 2002. p 1117–1122.
177. Burian J, et al. Report on the first patient group of the Phase I BNCT trial at the LVR-15 reactor. Sauerwein W, Moss R, Wittig A, editors. Bologna, Italy: Monduzzi Editore; 2002. p 1107–1112.
178. Coderre JA, et al. Tolerance of normal human brain to boron neutron capture therapy. Appl Radiat Isotopes 2004;61: 1084–1087.
179. Emami B, et al. Tolerance of normal tissue to therapeutic irradiation. Int J Radiat Oncol Biol Phys 1991;21:109–122.
180. Flickinger JC, et al. Development of a model to predict permanent symptomatic postradiosurgery injury for arteriovenous malformation. Arteriovenous Malformation Radiosurgery Study Group. Int J Radiat Oncol Biol Phys 2000;46:1143–1148.
181. Wittig A, Sauerwein WA, Coderre JA. Mechanisms of transport of *p*-borono-phenylalanine through the cell membrane in vitro. Radiat Res 2000;153:173–180.
182. Joel DD, et al. Effect of dose and infusion time on the delivery of *p*-boronophenylalanine for neutron capture therapy. J Neuro-Oncol 1999;41:213–221.
183. Morris GM, et al. Long-term infusions of p-boronophenylalanine for boron neutron capture therapy: evaluation using rat brain tumor and spinal cord models. Radiat Res 2002;158:743–752.
184. Smith DR, Chandra S, Coderre JA, Morrison GH. Ion microscopy imaging of ^{10}B from p-boronophenylalanine in a brain tumor model for boron neutron capture therapy. Cancer Res 1996;56:4302–4306.
185. Dahlström M, et al. Accumulation of boron in human malignant glioma cells *in vitro* is cell type dependent. J Neuro-Oncology 2004;68:199–205.
186. Bergenheim AT, Capala J, Roslin M, Henriksson R. Distribution of BPA and metabolic assessment in glioblastoma patients during BNCT treatment: a microdialysis study. J Neuro-Oncol 2005;71:287–293.
187. Henriksson R, et al. Boron neutron capture therapy (BNCT) for glioblastoma multiforme: A phase 2 study evaluating a prolonged high dose of boronophenylalanine (BPA) at the Studsvik facility in Sweden. Radiother Oncol (Submitted).
188. Mishima Y, et al. New thermal neutron capture therapy for malignant melanoma. Melanogenesis-seeking ^{10}B molecular-melanoma cell interaction from in vitro to first clinical trial. Pigment Cell Res 1989;2:226–234.
189. Hiratsuka J, Kono, Mishima Y. RBEs of thermal neutron capture therapy and $^{10}B(n,\alpha)^{7}Li$ reaction on melanoma-bearing hamsters. Pigment Cell Res 1989;2:352–355.

190. Tsuji M, Ichihashi M, Mishima Y. Selective affinity of ^{10}B-paraboronophenylalanine-HCI to malignant melanoma for thermal neutron capture therapy. Jpn J Dermatol 1983;93: 773–778.
191. Coderre JA, et al. Selective delivery of boron by the melanin precursor analog p-boronophenylalanine to tumors other than melanoma. Cancer Res 1990;50:138–141.
192. Mishima Y, et al. Treatment of malignant melanoma by single neutron capture therapy with melanoma-seeking ^{10}B-compound. Lancet 1989;1:388–389.
193. Mishima Y. Melanoma and nonmelanoma neutron capture therapy using gene therapy: overview. In: Larsson B, Crawford J and Weinreich, editors. Advances in Neutron Capture Therapy Vol. 1, Medicine and Physics. Elsevier; 1997. p 10–25.
194. Madoc-Jones H, et al. A phase-I dose-escalation trial of boron neutron capture therapy for subjects with metastatic subcutaneous melanoma of the extremities. In: Mishima Y, editor. Cancer Neutron Capture Therapy. New York and London: Plenum Press; 1996. p 707–716.
195. Yoshino K, et al. Improvement of solubility of p-boronophenylalanine by complex formation with monosaccharides. Strahlenther Onkol 1989;165:127–129.
196. Busse PM, et al. The Harvard-MIT BNCT Program: overview of the clinical trials and translational research. Proceedings of the 11th International Congress of Radiation Research, Vol 2. Dublin, Ireland, July 18–23, 1999. p 702–709.
197. Gonzalez SJ, et al. First BNCT treatment of a skin melanoma in Argentina: dosimetric analysis and clinical outcome. Appl Radiat Isotopes 2004;61:1101–1105.
198. Barth RF, et al. A nude rat model for neutron capture therapy of human intracerebral melanoma. Int J Radiat Oncol Biol Phys 1994;28:1079–1088.
199. Barth RF, et al. Neutron capture therapy of intracerebral melanoma: Enhanced survival and cure following blood-brain barrier opening to improve delivery of boronophenylalanine. Int J Radiat Oncol Biol Phys 2002;52:858–868.
200. Barth RF, et al. Boron neutron capture therapy of brain tumors: enhanced survival and cure following blood-brain barrier disruption and intracarotid injection of sodium borocaptate and boronophenylalanine. Int J Radiat Oncol Biol Phys 2000;47:209–218.
201. Kato I, et al. Effectiveness of BNCT for recurrent head and neck malignancies. Appl Radiat Isotopes 2004;61:1069–1073.
202. Pinelli T, et al. TAOrMINA: from the first idea to the application to the human liver. In: Research and Development in Neutron Capture Therapy. In: Sauerwein W, Moss R, Wittig A, editors. Bologna, Italy: Monduzzi Editore; 2002. p 1065–1072.
203. Ringe B, Pichlmayr R, Wittekind C, Tusch G. Surgical treatment of hepatocellular carcinoma: experience with liver resection and transplantation in 198 patients. World J Surg 1991;15:27085.
204. Iwatsuki S, et al. Hepatic resection versus transplantation for hepatocellular carcinoma. Ann Surg 1991;214:221–228.
205. Pinelli T. Neutron capture therapy for liver cancer metastases. Abstracts of the Eleventh World Congress on Neutron Capture therapy. Boston, October 11–15, 2004. p 52.
206. Suzuki M, et al. Biodistribution of ^{10}B in a rat liver tumor model following intra-arterial administration of sodium borocaptate (BSH)/degradable starch microspheres (DSM) emulsion. Appl Radiat Isotopes 2004;61:933–937.
207. Koivunoro H, et al. BNCT dose distribution in liver with epithermal D-D and D-T fusion-based neutron beams. Appl Radiat Isotopes 2004;61:853–859.
208. Chou FI, et al. Biological efficacy of BPA in malignant and normal liver cells. Abstract of the Eleventh World Cong on Neutron Capture Therapy. Boston, October 11–15, 2004. p 38.
209. Goodman JH, et al. Boron neutron capture therapy of brain tumors: biodistribution, pharmacokinetics, and radiation dosimetry of sodium borocaptate in glioma patients. Neurosurgery 2000;47:608–622.
210. Hideghéty K, et al. Tissue uptake of BSH in patients with glioblastoma in the EORTC 11961 phase I BNCT trial. J Neuro-Oncol 2003;62:145–156.
211. Coderre JA, et al. Biodistribution of boronophenylalanine in patients with glioblastoma multiforme: boron concentration correlates with tumor cellularity. Radiat Res 1998;149: 163–170.
212. Smith D, et al. Quantitative imaging and microlocalization of boron-10 in brain tumors and infiltrating tumor cells by SIMS ion microscopy: Relevance to neutron capture therapy. Cancer Res 2001;61:8179–8187.
213. Santa Cruz GA, Zamenhof RG. The microdosimetry of the ^{10}B reaction in boron neutron capture therapy: a new generalized theory. Radiat Res 2004;162:702–710.
214. Imahori Y, et al. Positron emission tomography-based boron neutron capture therapy using boronophenylalanine for high-grade gliomas: part 1. Clin Cancer Res 1998;4: 1825–1832.
215. Imahori Y, et al. Positron emission tomography-based boron neutron capture therapy using boronophenylalanine for high-grade gliomas: part 2. Clin Cancer Res 1998;4:1833–1841.
216. Takahasi Y, Imahori Y, Mineura K. Prognostic and therapeutic indicator of fluoroboronophenylalanine positron emission tomography in patients with gliomas. Clin Cancer Res 2003;9:5888–5895.
217. Kabalka GW, et al. The use of positron emission tomography to develop boron neutron capture therapy treatment plans for metastatic malignant melanoma. J Neuro-Oncol 2003;62: 187–195.
218. Bendel P. Biomedical applications of ^{10}B and ^{11}B NMR. NMR Biomed 2005;18:74–82.
219. Bendel P, Margalit R, Salomon Y. Optimized ^{1}H MRS and MRSI methods for the *in vivo* detection of boronophenylalanine. Magn Reson Med 2005;53:1166–1171.
220. Kajimoto Y, et al. Boron neutron capture therapy selectively destroys tumor cells preserving neurons in co-existing tumor lesion of malignant glioma. (Submitted)
221. Perry JR, et al. Challenges in the design and conduct of phase III brain tumor therapy trials. Neurology 1997; 49:912–917.
222. Shapiro W. Bias in uncontrolled brain tumor trials. Can J Neurol Sci 1997;24:269–270.
223. Stupp R, et al. Promising survival for patients with newly diagnosed glioblastoma multiforme treated with concomitant radiation plus temozolomide followed by adjuvant temozolomide. J Clin Oncol 2002;20:1375–1382.
224. Stupp R, et al. Radiotherapy plus concomitant and adjuvant temozolomide for gliomblastomas. N Engl J Med 2005;325: 987–996.
225. Agarwala SS, et al. Temozolomide for the treatment of brain metastases associated with metastatic melanoma: a phase II study. J Clin Oncol 2004;22:2101–2107.

See also IMMUNOTHERAPY; MONOCLONAL ANTIBODIES; RADIONUCLIDE PRODUCTION AND RADIOACTIVE DECAY; RADIOTHERAPY, HEAVY ION.

BRACHYTHERAPY, HIGH DOSAGE RATE

RUPAK DAS
University of Wisconsin
Madison, Wisconsin

INTRODUCTION

Brachytherapy is a form of radiotherapy whereby a radioactive source is used inside or at short distance from the tumor. There are three different forms of brachytherapy: interstitial, intracavitary, and skin therapy. In interstitial brachytherapy, the radioactive sources are implanted inside and throughout the tumor volume; in intracavitary brachytherapy the sources are placed in the body cavities very close to the tumor; while in skin therapy the sources are placed on the skin surface. Conventionally, brachytherapy implants have delivered the radiation at a low dose rate (dose rates of < 1 Gy $\cdot$ h^{-1}). Low dose-rate (LDR) interstitial implants can be temporary (meaning that the radioactive sources are left in place for a period of time, usually a few days, and then removed) or permanent (left in place without removal), while intracavitary implants are temporary. The advent of methods to deliver the dose at a much higher dose rates, in the range of 1–5 Gy $\cdot$ min^{-1}, brought an increase in the use of brachytherapy. All high dose-rate (HDR) brachytherapy treatments are temporary and treatments are administered using discrete fractions.

What Is a Remote Afterloader?

A remote afterloader (RAL) is a computer driven system that transports the radioactive source from a shielded safe into the applicator placed in the patient. Upon termination or interruption of the treatment, the source is driven back to its safe. The device may move the source by one of several methods, most commonly pneumatic air pressure or cable drives.

What is Stepping-Source Remote Afterloader?

A stepping source RAL is a particular design of the treatment unit that consists of a single source at the end of a cable that moves the source through applicators placed in the treated volume. The treatment unit can treat implants consisting of many needles or catheters in the patient. Multiple catheters are often required to cover the target with adequate radiation doses. Each catheter or part of an applicator is connected to the RAL through a channel. The computer drives the cable so that the source moves from the safe through a given channel to the programmed position in the applicator (dwell position) for a specific amount of time (dwell time). In any applicator, there may be many dwell positions. After treating all the positions in a given catheter (channel) the source is retracted to its safe and then driven to the next channel. The dwell positions and the dwell time in each channel are independently programmable, thereby giving a high level of flexibility of dose delivery. All currently available HDR RALs use the stepping-source design.

Currently there are three types of HDR RALs available in the market: MicroSelectron (Fig. 1, vendor Nucletron, Veenendaal, Netherlands), Gamma-Med (Fig. 2), and VariSource (Fig. 3, both marketed by Varian Associates, Palo Alto, CA).

The specific features of the three different RALs are shown in Table 1.

COMPONENTS OF A HIGH DOSE RATE REMOTE AFTERLOADER

While different in detail, all available HDR RALs consist of the same general components. Figure 4 gives an overview of the systems, with the major parts described below.

Shielded Safe

To provide a dose rate in the range of 1–5 Gy $\cdot$ min^{-1} in a RAL requires a ^{192}Ir source of 4–10 Ci. A shielded safe, which is an integrated part of the treatment unit, provides enough radiation shielding to house the source while not in treatment mode. Once in treatment mode, the source is driven out of the safe while it follows the program through the dwell positions. In the event of an interruption or termination of the treatment, the source is driven back to the shielded safe.

Radioactive Source

While delivering the HDR brachytherapy requires an intense source, passing the source through needles placed through a tumor requires one of a small size. The radioactive source in an HDR RAL is usually 3–10 mm in length

Figure 1. The Nucletron MicroSelectron V2 HDR RAL. The RAL wheels allow it to be conveniently positioned near the patient. The treatment head is mounted on a telescopic base that allows the head to be raised or lowered to the required height for treatment without moving the patient.

Figure 2. The Varian GammaMed RAL.

Figure 3. The Varian VariSource RAL.

and < 1 mm in diameter, fixed at the end of a steel cable (Figs. 5 and 6). The Nucletron source is placed in a stainless steel capsule and welded to the cable, while the Varian source is placed in a hole drilled into the cable and closed by welding. The ^{192}Ir radionuclide is now used for all HDR RALs, although early versions of HDR RAL used ^{60}Co. A new source has an activity near 10 Ci. Since ^{192}Ir has a half-life of 74 days, the source should be replaced every 3 months to keep the treatment in the HDR radiobiological regime. A trained medical physicist calibrates the source after each installation using a re-entrant well-type ionization chamber (Fig. 7). The chambers themselves are calibrated by secondary calibration laboratories known as Accredited Dosimetry Calibration Laboratories (ADCL). The resulting source calibration is verified against the manufacturer's source calibration.

Source Drive Mechanism

When the RAL unit receives a command to initiate a treatment, the stepper motor connected to the reel containing the drive cable turns, causing the source cable to advances from the shielded safe along a path constrained by transfer tubes to the first treated dwell position in the applicator attached to the first channel. The source dwells at that position for a predetermined duration (dwell time) as calculated by the treatment planning system (see below). After completing that dwell, it goes on to the subsequent dwell positions. Some units step as the source drives out (MicroSelectron), stopping first at the dwell position most proximal to the afterloader, while the other (VariSource and Gamma-Med) the source travels first to the most distal dwell (toward the tip of the applicator), and a bit farther, and then steps as the source returns toward the safe. Stepping on the outward drive obviates any concern about the effect of slack in the drive mechanism affecting the accuracy of the source position. The unit that steps on the way back into the unit includes correction for slack in the calibration of the source location. Upon completion of the treatment for the first channel, the source is retracted into the safe, and redirected to travel to the second channel. The process is repeated for all the subsequent treatment channels. The programmed movement of the source is verified by means of an optical encoder or other devices that compare the angular rotation of the stepper motor or cable length ejected or retracted with the number of pulses sent to the drive motor. This system is capable of detecting catheter obstruction or constriction as increased friction in the cable movement. Under certain fault conditions, if the stepper motor fails to retract the source, a high torque direct current (dc) emergency motor will retract the source.

The confirmation of the source exit from and return to the safe is carried out by an "optopair", consisting of a pair of light-sensitive detector and infrared (IR) light source, that detects the cable when its tip obstructs the light path. All the currently marketed after-loaders are also equipped with check cables or dummy sources. The check cable is an exact duplicate of the radioactive source along with its cable, except not radioactive. Before the ejection of the radioactive source, the check cable is first ejected to check the integrity of the catheter system. After a noneventful check by this "dry run" with the dummy source, the radioactive source is then sent for treatment.

Indexer

The RALs are equipped with an indexer, shown in Fig. 8. The indexer consists of an S-tube (item 14 in Fig. 4) that directs the source cable from the exit of the safe to one of the exit ports from the unit (channels). The various catheters or applicator parts connect to these channels, usually through connecting guides called transfer tubes. Different units have between 3 and 24 channels available for

Table 1. Specific Features of the Three Currently Marketed HDR RALs

	MicroSelectron V2	Gamma Med+	VariSource 200/200t
Vendor	Nucletron	Varian	Varian
Sources	10 Ci of ^{192}Ir	10 Ci of ^{192}Ir	10 Ci of ^{192}Ir
Source dimension	3.5 mm L, 1.1 mm OD	4.52 mm L, 0.9 mm OD	5 mm L, 0.59 mm OD
Channels	18	3 or 24	20
Source extension	1500 mm	1300 mm	1500 mm
Channel length	Variable	Fixed	Variable
Source movement	Stepping forward	Stepping backward	Stepping backward
Step sizes	2.5, 5 or 10 mm	1–10 mm, 1 mm steps	2–99 mm, 1 mm steps
Dwells/channel	48	60	20
Speed of source	50 cm · s^{-1}	60 cm · s^{-1}	50–60 cm · s^{-1}

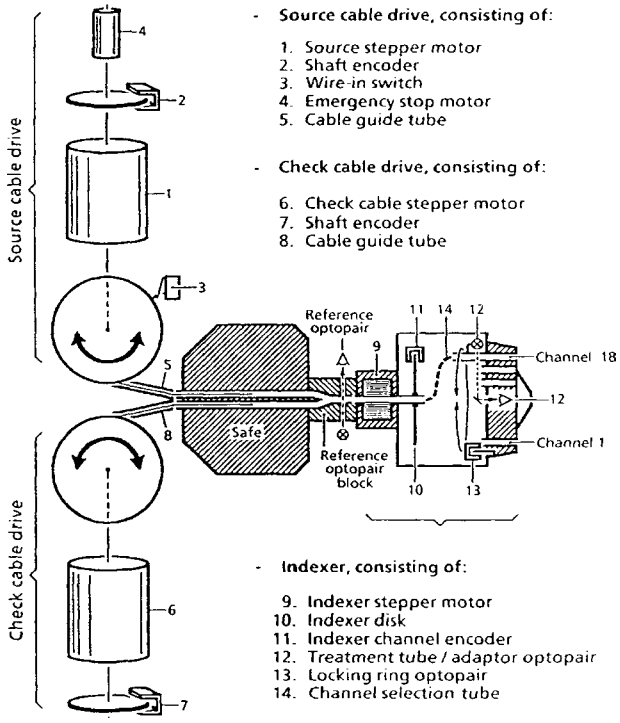


Figure 4. Schematic diagram of a single stepping source RAL. (Courtesy of Nucletron Corporation, Columbia, MD.)

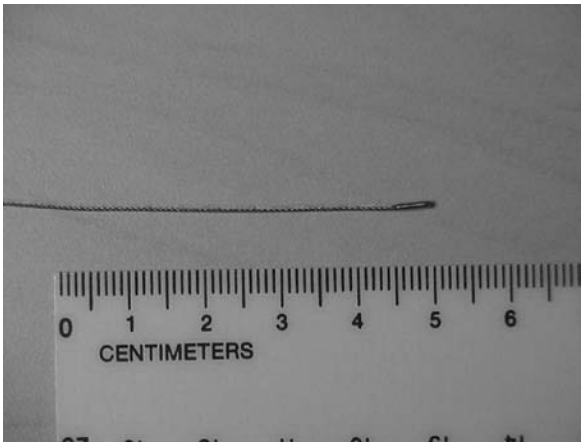


Figure 6. A ^{192}Ir HDR source for the MicroSelectron at the end of a steel drive cable, as shown in Fig. 5.

connection. If a patient's treatment requires more than the number of channels on a given treatment unit, the treatment must be broken into sessions, where the catheters are connected up to the number of channels available and treated Then the transfer tubes are disconnected from the catheters just treated and reconnected to the next set of catheters for continuation of the treatment.

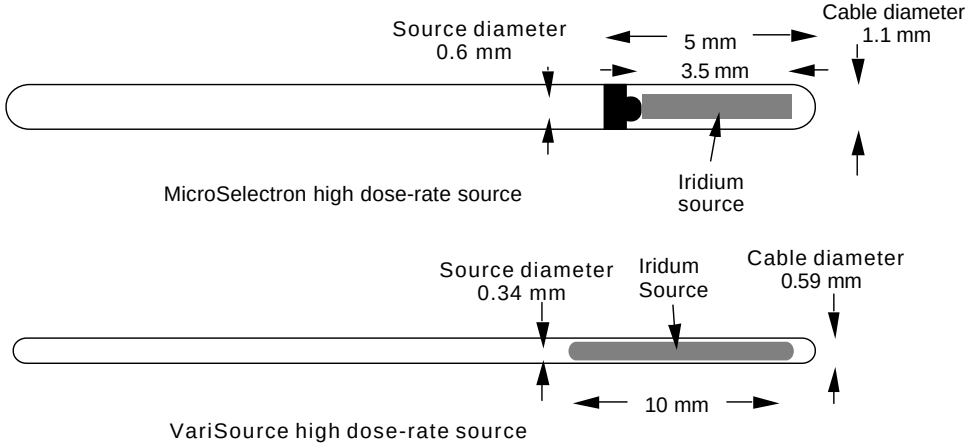


Figure 5. Schematics of the two types of sources used in stepping-source RALs. The VariSource is an earlier version, while the new source has a length of 5 mm.

Figure 7. A re-entrant well-type ionization chamber used for calibration of the HDR brachytherapy sources.

Transfer Tubes

Transfer or guide tubes are long tubes that act as a conduit to transfer the source from the RAL to the applicators or catheters for treatment. One end of the transfer tube is attached to the indexer of the RAL (Fig. 9), while the other end is attached to the interstitial, intracavitary, or transluminal applicators (Fig. 10). The applicator-end of the transfer tube contains spring-loaded ball bearings that block the path through the tube if no applicator is attached. When an applicator is inserted, it pushes aside the ball bearings, opening the path for the source cable. When the

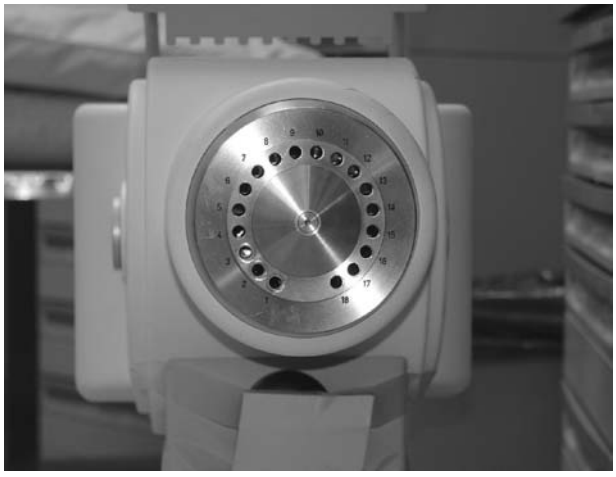


Figure 8. The frontal view of an indexer from the Nucletron, MicroSelectron HDR RAL, consisting of 18 channels.

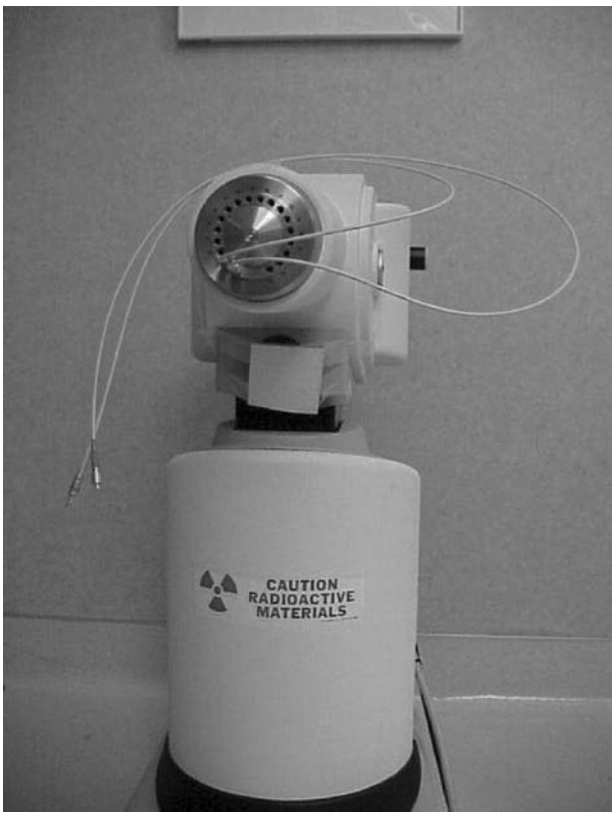


Figure 9. A view of two types of transfer tubes hooked up to the indexer of a RAL.

Figure 10. View of the transfer tubes connected to a gynecological applicator. Ball bearings beneath the gray polymer coating allow verification of the proper connection of the transfer tubes to the applicator. The number 1 and 2 represents that these transfer tubes must connect the channel 1 and 2 of the indexer ring and the similarly numbered parts of the applicator.

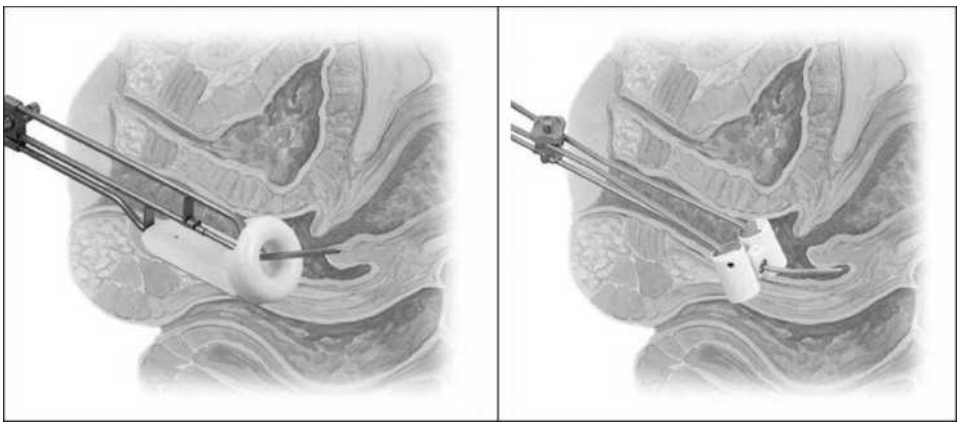


Figure 11. Gynecological applicator used for the treatment of cervical cancer. (Courtesy of Nucletron Corporation, Columbia, MD.)

check cable makes its test run, if no applicator is attached to the transfer tube, the check cable hits the obstacle of the ball bearings, and prevents ejection of the source. Each type of applicator has its own type of transfer tube.

Applicators

An array of applicators for different treatment sites are marketed by each vendor. Each vendor designs their own applicators that can only be used with their transfer tubes and HDR RALs. Figure 11 shows two cervical applicators marketed by Nucletron used for the treatment of cervical cancer.

Treatment Control Station

The treatment control station (Fig. 12) allows the user to select the source travel and dwell sequence to be used in each channel. This can be entered by three ways: (*1*) manually by the keyboard/mouse at the control station; (*2*) recalling a standard plan from the computer and then editing the data without affecting the standard plan from which it originated; or (*3*) by importing the data from a treatment planning system via transfer medium or a network connection to the treatment control station.

Treatment Control Panel

The treatment control station transfers the data to the treatment control panel. A hard or soft START button initiates the execution of the treatment according to the program. In addition, there is an INTERRUPT button, which when pressed retracts the source and stops the timer, allowing the user to enter the treatment room without receiving radiation exposure. A RESUME or START button resumes the treatment from the time and the dwell position where it was interrupted. A master EMERGENCYOFF button initiates the high torque dc emergency motor to retract the source. In the normal course of a successful termination of the treatment, the timer runs to zero and the machine automatically retracts the source. Figure 13 shows an example of the treatment control panel.

SAFETY FEATURES

The HDR RALs are complicated devices containing very high activity radioactive sources. Serious accidents can

Figure 12. A view of the monitor of the RAL treatment control station.

Figure 13. The Treatment Control Panel of the Nucletron, microSelectron HDR RAL. The START button is the white button on the right side of the panel, while the EMERGENCY OFF button is the top button on the left side of the panel.

Figure 14. A view of the access panel of the MicroSelectron treatment unit. The center button is an emergency stop button on the treatment unit. Also showing are the manual retraction of the radioactive source cable (left) and the check cable (right).

happen quickly. All such units have many safety features and operational interlocks to prevent errant source movement or facilitate rapid operator response in the event of a system failure.

Emergency Switches

Numerous EMERGENCY OFF switches are located at convenient places and are easily accessible, in case a situation arises. One EMERGENCY OFF switch is located on the control panel. Another EMERGENCY OFF button is located on the top of the remote afterloader treatment head. Vendors usually install one or two emergency switches in the walls of the treatment room. In the event a treatment is initiated with someone other than the patient in the treatment room, that person can stop the treatment and retract the source by pressing the EMERGENCY OFF button. Figure 14 shows the EMERGENCY OFF switch on the treatment unit.

Emergency Crank

All HDR RALs have emergency cranks to retract the source cable if the source fails to retract normally and the emergency motor also fails to reel in the source. Figure 14 shows such a crank for the MicroSelectron and Fig. 15 for the VariSource. Using the crank requires the operator to enter the room with the source unshielded. Exposure rates for this situation are considered below.

Door Interlock

Interlock switches prevent initiation of a treatment with the door open. While in progress, opening the door interrupts the treatment. This safety feature protects the medical personnel from radiation exposure, in the event somebody enters the treatment room without the knowledge of the operator. If a door is inadvertently opened during the treatment, the treatment is interrupted and

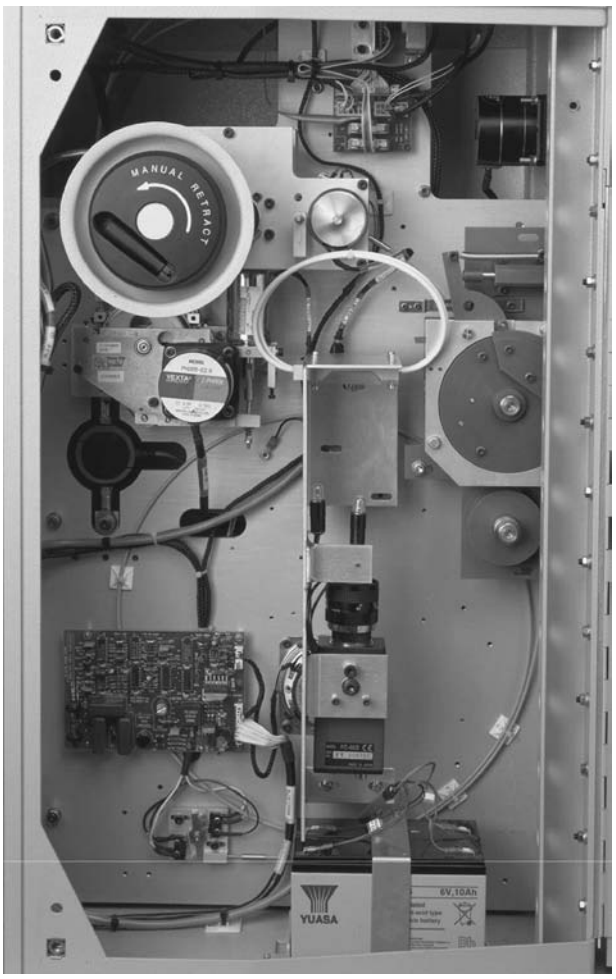


Figure 15. The back panel of the VariSource showing the crank for manual source retraction in an emergency.

the source returns to the safe. The treatment can be resumed at the same point where it was interrupted by closing the door and pressing the START or the RESUME button at the control panel.

Audio–Visual System

All HDR brachytherapy suites are equipped with a closed circuit television system (CCTV) or shielded windows and/or mirrors for observing the patient, and a two-way audio system to communicate with the patient during treatment.

Radiation Monitor and Treatment on Indicator

Three separate independent systems alert personnel when the source is not shielded. One radiation detector is part of the treatment unit and indicates on the control panel when it detects radiation. An independent unit, usually mounted on the treatment room wall with displays both inside and outside the room also alerts the operator and other personnel when the radioactive source is out of the safe. A TREATMENT ON Indicator outside the room, activated when the source passes the reference optical pair discussed above and shown in Fig. 4, also indicates that a treatment is in progress.

Table 2. Exposure Rates from an Exposed 10 Ci ^{192}Ir Source

Typical Situation	Distance, m	Dose Equiv Rates, Sv · h^{-1}	Time, min, to Receive	
			10 Gy (likely injury)	0.5 Sv (annual body limit)
In Patient	0.01	460	1.25 min	0.07 min
Handling with Kelly Clamps, to hands	0.1	4.6	2.1 h	6.5 min 98 min for hand limit
Handling with Kelly Clamps, to body	0.3	0.5	18.8 h	1.0 h
Standing near	1	0.046	8.7 days	11 h
Standing far	2	0.012	34.8 days	43 h

Emergency Service Instruments

In the event the radioactive source fails to retract after termination, interruption, pushing the EMERGENCY SWITCH, or cranking the stepper motor manually, the immediate priority is to remove the source from the patient. Table 2 gives the exposure rates at various distances from a 10 Ci ^{192}Ir source. Table 2 shows that the dose to the patient, with the source in contact, can cause injury in a very short time. On the other hand, the operator, working at a greater distance, is unlikely to receive a dose exceeding regulatory limits for a year, let alone one that would cause health problems. Once the source is removed from the patient and moved to a distance of even a meter, the exposure rate is quite low, and whatever actions need be taken to remove the patient from the room can be performed safely.

The effective annual limit to the body should actually be 10 times < the 0.5 Sv in keeping with the principle to keep exposures as low as reasonable achievable (ALARA), and ideally should not be received in one, short exposure. The allowed exposure to the hands is 15 times that to the body.

The preferred approach to a source that will not retract by any of the methods is to remove the applicator from the patient as quickly as possible, and place the applicator containing the source in a shielded container (Fig. 16). If it is clear that the cable is caught in the transfer tube and not in the applicator itself, the applicator or catheter may be disconnected from the transfer tube and the source pulled from the applicator. In some cases, this will be faster than removing the applicator. The reason to avoid disconnecting the applicator from the transfer tube is that a source may stay in the applicator if the source capsule shatters. In that case, removing the applicator attached to the transfer tube keeps the system closed, while disconnecting the two opens a path for parts of a broken source to fall from the applicator into body cavities or crevices, or roll onto the floor.

A situation may arise when the source needs to be detached manually from the treatment unit. One (still unlikely) scenario would be if the source were stuck out of the treatment unit, the sources or the closed applicator had been removed from the patient, a person were pinned very close to the source so neither they, nor the treatment unit, could be moved, and the source on the cable could not reach the shielded container. In this special situation, the source cable should be cut from the unit and the source placed in the shielded container always present in the room. In cutting the source cable, it must be clear that the cut is *not* through the source capsule. For units with the capsule welded on the cable, the cut must be through the braided cable as opposed to the smooth steel capsule (Fig. 17). For sources imbedded in the cable, a sufficient length of the cable must be seen to assure the cut occurs behind the source. Thus, emergency tools that must be present in the treatment room and always readily accessible include a wire cutter, a pair of forceps, and a shielded service container.

Back-Up Battery

In case of a power failure during the treatment, the machine is equipped with a back-up battery to provide retraction of the source to its safe. The batteries should be tested with each source exchange.

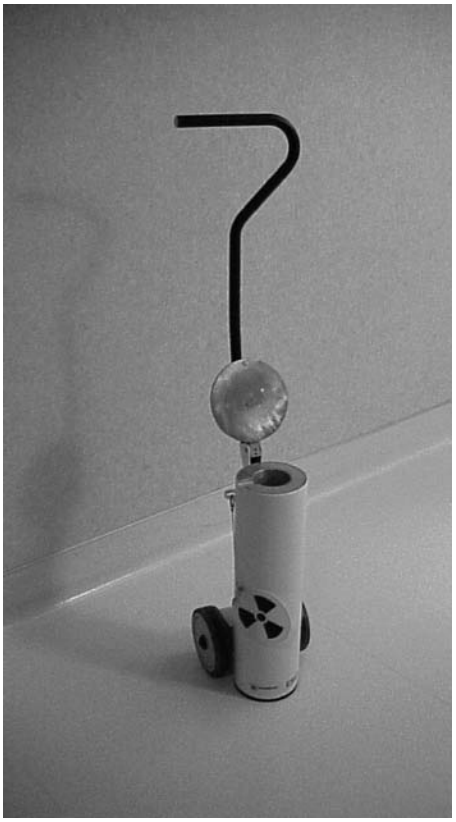


Figure 16. The shielded container for emergency placement of an unretracted source.

Figure 17. Cutting the source cable from a treatment unit. This procedure should only be performed in very special, rare situations as described in the text. Great care must be taken to assure the cut is through the cable and not the source capsule.

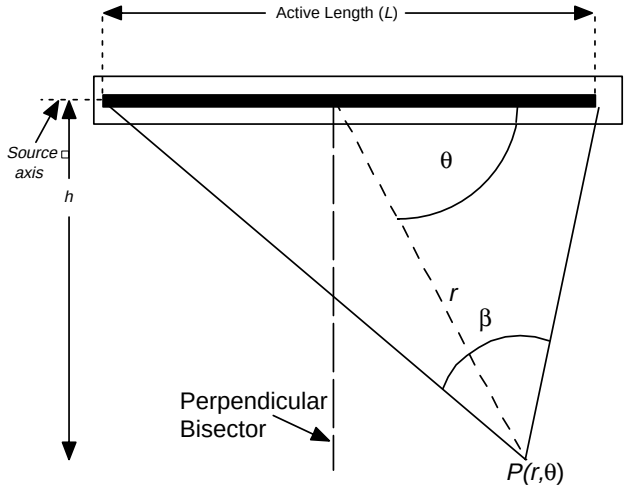


Figure 18. Geometry and legend for the dosimetry of a line source as given in Eq. 1.

TREATMENT PLANNING SYSTEM

Software and hardware for the treatment planning system are provided by the vendor selling the treatment unit. Three-dimensional (3D) patient data [computed tomography (CT), magnetic resonance imaging (MRI)] can be directly transported and loaded in the planning system. Two dimensional (2D) data (e.g., from radiographs) usually are loaded interactively by computer peripherals (scanners, digitizers) although some automated input systems are available. With 2D input, the target information must be inferred since tumors are generally not visible on the images, while the 3D imaging often visualizes tumors as well as surrounding normal tissue structures. With either input, tumor volume is entered on these images, and the treatment-planning volume is constructed by adding some margin to the tumor volume. Various computer algorithms help the planner conform the prescribed dose to the target volume. Data characteristic for the radioactive source are usually supplied by the vendor and included in the software. The medical physicist enters the source strength data both in the planning system and the treatment unit at the time of the installation of the new source in the treatment unit after calibration.

Dose Calculation

The treatment-planning computer calculates the dose distribution for a patient containing an applicator with a given set of dwell positions, each with their own dwell time. In calculating the dose distribution, the computer first calculates the dose to a set of grid points. Usually, the operator wishes to see the results presented as isodose lines. An isodose line on a given plan connects all points receiving the same dose, much like elevation lines on a contour map of part of the Earth connects points with the same altitude. From the dose values at the grid points, the computer interpolates to find the path of the isodose line value specified by the user. The calculation of the dose from one dwell position, identified with the subscript i, to a point $P(r,)$ as shown in Fig. 18, uses the formula (1),

$$D_i(r_i,\theta) = S_K \cdot \Lambda \cdot [G_i(r,\theta)/G(r_o,\theta_o)] \cdot g_i(r_i) \cdot F_i(r_i,\theta) \cdot t_i \quad (1)$$

where

$D_i(r_i,\theta)$ = The radiation dose to water, in units Gy, at position $P(r,\theta)$

S_K = The air kerma strength of the source in $\mu Gy m^2 \cdot h^{-1}$. The strength of photon-emitting brachytherapy sources usually is specified by the intensity of the radiation at some distant outside of the source rather than by the amount of radioactivity contained inside. In this manner, variations in the source encapsulation, which may attenuate varying amounts of the radiation given off by the contained radionuclide, have no effect on the dose delivered to the patient. While the strength of a new HDR source is often quoted as 10 Ci (the approximate activity in the capsule), the actual source strength determination accounts for the energy of radiation from the source, transferred to a mass of air at a given point. Air kerma is the energy transferred from the radiation to kinetic energy in the medium per unit mass, where the medium must be specified. That point for air kerma strength is at 1 m, and a new source would have an air kerma strength of ~ 40 mGy $\cdot$ m$^2 \cdot$ h^{-1}, or in shorthand, 40 kU, where 1 U $= \mu$Gy $\cdot$ m$^2 \cdot$ h^{-1}. A radiation dose of 1 gray (Gy), equals 1 joule per kilogram (1 J $\cdot$ kg).

Λ = The dose rate constant, that is the absorbed dose rate in cGy $\cdot$ h^{-1} at 1 cm from the source in the perpendicular plane that bisects the source axis per unit air kerma strength. For ^{192}Ir sources, $\Lambda = 1.12$ cGy $\cdot$ h^{-1}/U.

$[G_i(r_i,\theta_i)/G(r_o,\theta_o)]$ = The geometry function, which accounts for changes in dose rate due to the relative positions of the source and the calculation point and the

shape of the source. The numerator expresses the geometric dose pattern for the point of calculation while the denominator gives that for the reference condition, where $r_o = 1$ cm and $\theta_o = 90^\circ$. The geometric dose pattern usually is approximated as $1/r^2$ for a point source, and for a line source /(L·h) as shown in Fig. 18.

$g_i(r_i)$ = The radial dose function, variation in the dose rate with distance from the source due to the attenuation and scatter due to the tissue between the source and the point of calculation at distance r, normalized at 1 cm, and not including any effect in dose rate due to geometry (i.e., the geometric function has been removed from the dose at the calculation distance and at 1 cm for the ratio).

$F_i(r_i,\theta)$ = The anisotropy function, which describes the deviation of the shaped of the isodose lines from a circle. The function $F_i(r_i,\theta)$ = the dose at the calculation point $P(r,\theta)$ divided by the dose at the same distance, r_i but on the perpendicular bisector, that is with $\theta_o = 90^\circ$, and, as with the radial dose function, with the geometrical effects removed.

t_i = the dwell time for dwell position i.

Dose Optimization

The treatment planning addresses first which dwell positions will be used. Then for each of the dwell positions, the dwell time must be calculated. The goal is to match the resulting dose distribution with the target volume, a process referred to as optimization. There are several methods to assist the operator in optimizing the dwell times. The methods fall into three main categories:

Analytic methods use relatively simple algorithms to calculate the dwell time for each dwell position. One of the most common, geometric optimization (2,3), weights the dwell time for a position inversely to the sum of the doses to that position from the other dwell positions. For an intracavitary application, where the dose distribution is intended to more or less conform to the shape of the applicator tracks, "distance optimization" is used, where the contributions of all of the other dwell positions are included in the summation of the dose. For interstitial application, where the implant usually treats a volume, the process becomes "volume optimization", with the dose contributions from dwell position along the same track excluded in the summation.

Dose specification methods attempts to calculate the dwell times to deliver a specified dose to designated points (dose points or optimization points) placed throughout the volume or on the surface of the target (4–6). The dose to each specified point described an equation with the dose to the point on one side and the contributions from each of the dwell positions on the other,

$$\text{Dose} = S_K \cdot \Lambda \cdot \sum_{i}^{\text{all dwells}} \{[G_i(r_i,\theta)/G(r_o,\theta_o)] \cdot g_i(r_i) \cdot f_i(r_i,\theta) \cdot t_i\} \quad (2)$$

In the simplest situation, such an approach becomes solving a set of simultaneous equations for the doses to the optimization points for the unknown dwell times. The problem comes when there are more equations than unknowns (more dose points than dwell positions and the set is over determined) or the converse (when the set is underdetermined). In the general case, the set is solved by a least-squares method to find the values for the dwell times that minimizes the square of the difference between the doses desired and the doses resulting from times in the set of equations. However, it is very helpful to add an additional criterion on the dwell times: controlling the fluctuation in dwell times between adjacent positions. The optimization equation becomes

$$X^2 = \sum_{j=1}^{\text{All dose points}} (D_j^{\text{precribed}} - D_j^{\text{calculated}})^2 + V \sum_{i=1}^{\text{dwells}-1} (t_{i+1} - t_i)^2 + u \sum_{i=1}^{\text{all dwells}} t_i \quad (3)$$

where the first term considers the difference between the prescribed dose and that calculated dose for each specified point. The value of the second term depends on the differences in dwell times between each dwell position and its neighbor. The factor, the dwell weight gradient factor, v, determines how important minimizing this fluctuation is. Large values for v (> 1) tend to force the dwell times to be the same, not producing a very conformal dose distribution; small values (< 0.2) permit negative times to result from the optimization (not a physical situation). The last term minimizes the overall exposure time, assuming that the set of dwell times that adequately treats the target volume with the least total time results in the lowest dose to the rest of the patient. The factor u determines how important minimizing dwell time is for the optimization. The optimized set of dwell times minimizes X^2.

Stochastic methods use iterative techniques to find adequate values for the dwell times. Generally, these approaches establish an objective function, such as the difference in dose to a set of point between that prescribed and that achieved. The function may also include penalties for excessive doses to normal tissue structures or lack of homogeneity through an implanted volume. The goal of the optimization is to obtain the best score for the objective function. The process begins with a set of values for the dwell times, evaluates the objective function, and then makes changes in the dwell times. If the new set of times improves the value of the objective function, the new set becomes the current best set. If the old set of times gave a better value for the objective function, it remains the current best set. Obviously, the strategy for how to pick each new set of values in the core of the methodology, and a more complete discussion is beyond the scope of this text. The most common approaches in the literature are simulated annealing (7) and the genetic algorithm (8).

The goal of all the optimization methods is to adequately treat the target tissues while sparing the sensitive normal structures. A resultant plan is shown in Fig. 19.

SHIELDING

The radioactive source in the high dose rate machine starts ~ 10 Ci with an exposure rate at a distance of 1 m from the

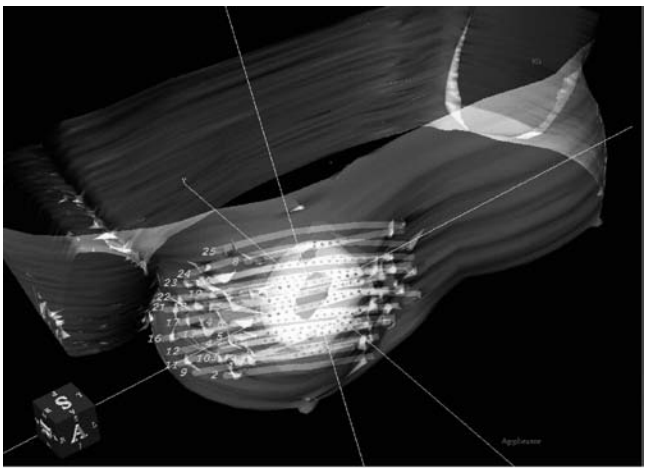


Figure 19. A 3D view of dose distribution of a breast catheter implant with 25 catheters. The inner volume is the tumor volume (lumpectomy cavity), while the outer volume (planning treatment volume) has been generated by adding some margins to the tumor volume. The gray cloud is the dose distribution generated by the treatment planning system.

source of ~ 46 mSv $\cdot$ h^{-1}. According to the rules and regulations of the United States Nuclear Regulatory Commission (USNRC), the annual limit for radiation exposure to the public is 1 mSv and the annual occupational limit is 5 mSv. (The actual limit for occupationally exposure persons is 50 mSv $\cdot$ year^{-1}, but following the principle of maintaining exposures as low as reasonable achievable, the NRC usually holds licensees to exposure 0.1 of the limit.) In addition to the annual limit, NRC requires that in an unrestricted area the dose equivalent rate should not be > 0.02 mSv in any given hour. Thus, the HDR machine needs to be housed in an adequately shielded room. To meet these requirements in a HDR brachytherapy suite, where the walls and the ceiling are at least 1.5 m from the machine head, concrete wall of ~ 43–50 cm (or 4–5 cm of lead) are needed. For larger rooms the concrete wall thickness will be lower since the exposure rate is inversely proportional to the square of the distance from the radioactive source. For details on the procedures for calculating the thickness of barriers for a particular facility, see health physics texts such as Cember (9) or McGinley (10).

QUALITY ASSURANCE

In order to maintain the quality of patient care, a quality management program is required in every facility that provides HDR brachytherapy treatment. Such a program generally follows standards set by professional organizations and intends to minimize untoward events caused by the malfunction of the machine or human error. Such programs become exceedingly important in HDR brachytherapy because the planning and the treatments tend to happen very quickly, increasing the likelihood of accidents and mistakes. Quality Assurance (QA) tests measure some performance aspect of the treatment unit and compare the results with expectations in order to demonstrate proper operation. QA is performed at various intervals: some for each patient, some once each treatment day, and others with each source change. Moreover, for HDR RAL, the USNRC mandates that users meet certain standards, including education and training on operating the machine, emergency procedures, radiation monitoring, pre-treatment safety checks, safe and accurate delivery of the treatment, and monthly/initial calibration of the source. Since the details of quality assurance is outside the scope of this literature, interested readers can refer to the report of Task Groups 59 (11) and 56 (12) of the American Association of Physicists in Medicine and relevant texts (13).

In general, the problem of quality assurance becomes assuring that the treatment will deliver the correct dose, to the correct location, safely. Thus, the tests generally follow the outline below:

Verification of dose variables.

- Checking the strength of the source compared with that projected from the initial calibration based on radioactive decay.
- Checking the proper operation of the controlling timer.

Verification of position control.

- Checking that the source goes to the location programmed.
- Checking coincidence between the programmed positions and the respective positions indicated by imaging markers.
- Checking consistent movement of the source.

Verification of proper operation of safety features.

- Checking operation of the door interlocks.
- Checking the operation of a handheld radiation detector.
- Checking the operation of the on-board and on-wall radiation detectors.
- Checking the operation of the check cable runs and interlocks.
- Checking the operation of the EMERGENCY OFF and TREATMENTINTERRUPT buttons.

COSTS

Currently, two vendors (Nucletron Corporation and Varian Medical Systems) market their RAL treatment unit in the United States. Both the devices requires a capital expenditure of $\sim$\$500,000–750,000, which includes the treatment unit, a variety of transfer tubes, along with the software and hardware for the treatment planning system. Applicators that are needed to be placed in the tumor costs extra. The costs of preparing a shielded room along with ancillary equipment for X-ray imaging and operating room

procedures can be another $500,000–750,000. Hence, the total cost can run in between $1–1.5 million (14).

ADVANTAGES AND DISADVANTAGES

HDR comparing with LDR brachytherapy offers several advantages and disadvantages. Being aware of these permits safe and effective operation and application of HDR brachytherapy.

Advantages of HDR Brachytherapy

Safety. One of the major advantages of a RAL is the reduction or elimination of radiation exposure to the radiotherapy staff. In conventional LDR, manual afterloading, the radiotherapy staff receives radiation exposure while loading the applicators with the radioactive sources, and the nursing personnel are exposed during patient care through the duration of the treatment (1–4 days). With either HDR or LDR remote afterloading, the radiotherapy personnel are outside the shielded room during the treatment, and hence are exposed to minimal radiation.

Optimization. The design of the HDR RAL with the stepping source allows greater flexibility and control over dose distribution. The stepping source allows optimization of the dose distribution by adjustment of the dwell times for each dwell position in each channel. The dwell times can be varied infinitely, permitting very fine control of the dose distribution. In LDR, either manual or using an RAL, the finite number of activities available (usually four at most) and the larger sources used with manual applications impose a restriction on the ability to conform the dose distribution to the target.

Stability. Because HDR intracavitary treatments take so little time ($\sim$ 1 h), applicators can be fixed in place much more stably than for the several day treatments using LDR brachytherapy.

Dose Reduction to Normal Tissue. As with stability, the short duration of HDR intracavitary treatments allows displacement of normal tissue structure (i.e., pushing them away from the source paths) to a greater extent than with LDR treatment.

Applicator Size. The small size of the HDR source permits the use of smaller applicators than those required for the LDR applications, increasing the comfort to the patient.

Outpatient Treatment. Almost all HDR patients are treated on an outpatient basis compared to LDR patients who usually are treated as inpatients. Outpatient treatment is more convenient for the patients and generally results in lower overall costs.

Disadvantages of HDR Brachytherapy

Investment. The initial expense of HDR RAL is very high. Machines and site preparation costs can be anywhere between $0.5 and 1 M.

Complexity. The technological complexity of HDR RAL opens the increased probability of errors, and leads to increased regulatory scrutiny.

Compressed Time Frame. As mentioned above, the rapidity with which procedures progress in HDR brachytherapy increases the probability of executing errors.

Radiobiology. As the dose rate increases, the radiosensitivity (damage per unit dose) increases for both normal tissues and tumors. Unfortunately, the radiosensitivity for the normal tissue increases faster than that for tumors, increasing the likelihood of injuring the patient while controlling the tumor. Overcoming this radiobiological handicap requires the use of the advantages of *optimization*, *stability*, and *dose reduction to normal tissues*, in addition to fractionization. As with external-beam radiotherapy delivered using a linear accelerator, which also operates in the HDR regime, spreading the treatments over many smaller fraction delivered over several days reduces the difference in radiosensitivities between the tumor and the normal tissues.

BIBLIOGRAPHY

Cited References

1. Nath R, Anderson LL, Luxton G, Weaver KA, Williamson JF, Meigooni AS. Dosimetry of interstitial brachytherapy sources: Recommendations of the AAPM Radiation Therapy Committee Task Group No. 43. Med Phys 1995;22:209–234.
2. Edmundson GK. Geometry based optimization for stepping source implants. Activity—The Selectron User's Newsletter 1991;5:22.
3. Edmundson GK. Geometric optimisation: an American view. In: Mould RF, editor. International brachytherapy. Veenendaal, The Netherlands: Nucletron International BV; 1992, p 256.
4. van der Laarse R. Optimization of high dose rate brachytherapy. Activity—The Selectron User's Newsletter 1989;2:14–15.
5. van der Laarse R, Edmundson GK, Luthmann RW, Prins TPE. Optimization of HDR brachytherapy dose distributions. Activity—The Selectron User's Newsletter 1991;5:94–101.
6. van der Laarse, Thomadsen BR, Houdek PV, van der Laarse R, Edmunson G, Kolkman-Deurloo I-KK. Treatment planning and optimization. In: Nag S, editor. High dose rate brachytherapy: a textbook. Armonk, (NY): Futura Publishing Co.,; 1994. p 85–91.
7. Sloboda RS. Optimization of brachytherapy dose distributions by simulated annealing. Med Phys 1992;19:955–964. See also, Sloboda RS, Pearcey RG, Gillan SJ. Optimized low dose rate pellet configuration for intravaginal brachytherapy. Int J Radiation Oncol Biol Phys 1993;26:499–511.
8. Davis L. Handbook of Genetic Algorithms. New York: Van Nostrand Reinhold; 1991.
9. Cember H. Introduction to health Physics. 3rd ed. New York: McGraw-Hill; 1996.
10. McGinley P. Shielding Techniques for Radiation Oncology Facilities 2nd. Madison: Medical Physics Publishing; 2002.
11. Kubo HD, Glasgow GP, Pethel TD, et al. High dose-rate brachytherapy treatment delivery: AAPM Radiation Therapy Committee Task Group No. 59. Med Phys 1998;25:375–403.

12. Nath R, Anderson LL, Meli JA, et al.Code of practice for brachytherapy physics: AAPM Radiation Therapy Committee Task Group No. 56. Med Phys 1997;24:1557–1598.
13. Thomadsen BR. Achieving Quality in Brachytherapy. Bristol: Institute of Physics Press; 1999.
14. Rivard MJ, Kirk BL, Stapleford LJ, Wazer DE. A comparison of the expected costs of high dose rate brachytherapy using ^{252}Cf versus ^{192}Ir. Appl Radiat Isot 2004;61:1211–1216.

See also BRACHYTHERAPY, INTRAVASCULAR; HYPERTHERMIA, INTERSTITIAL; PROSTATE SEED IMPLANTS; RADIATION DOSIMETRY FOR ONCOLOGY.

BRACHYTHERAPY, INTRAVASCULAR

FIRAS MOURTADA
MD Anderson Cancer Center
Houston, Texas

INTRODUCTION

Intravascular brachytherapy (IVB) is a novel treatment modality that delivers ionizing radiation to a coronary artery to prevent renarrowing, that is, restenosis caused by stent placement within the artery. The term *brachy* is the Greek word for near since the radioactive source is placed inside or near the target cells. In general, the field of brachytherapy has been practiced for decades in the field of radiation oncology for treatment of intracavitary (vagina, bronchus, esophagus, rectum, nasopharynx, etc.) and interstitial (muscle sarcoma, prostate, breast, etc.) cancers. As a subspecialty, IVB is relatively new where most of its development took place in the 1990s. Ionizing radiation describes both electromagnetic (γ rays, X rays) and particulate (neutrons, beta, and alpha particles) of sufficient energy to remove electrons from the target atom (thus ionizing). Unlike conventional brachytherapy where the target is mostly centimeters away from the source, IVB targets the adventitia of the vessel wall, located within 1–5 mm from the radioactive source. To obtain accurate dosimetry data in such close range is a challenge. The scope of this article is on the delivery devices for IVB and tools needed to assess the dosimetric properties of such brachytherapy devices. (Dosimetry is a subspecialty of radiation physics that deals with the measurement of the absorbed dose or dose rate resulting from the interaction of ionizing radiation with matter.) Such techniques are useful and can be applied in other future applications that require delivery of ionizing radiation to a target within a few millimeters from a radioactive source.

Mechanisms of Restenosis

A diseased coronary vessel is mainly caused by atherosclerotic plaque formation containing mostly cholesterol and lipids. This condition can lead to a heart attack and chest pain (angina) where the blood flood within the lumen is compromised. Coronary artery bypass surgery (CABG) is the traditional method to alleviate this condition. In the last few decades, minimally invasive procedures have been developed in a field known as *Interventional Cardiology*. Percutaneous transluminal coronary angioplasty (PTCA), first performed by Gruentzig in 1977 (1); and endovascular prosthetic devices (stents), first performed by Dotter et al. (2) and Cragg et al. (3) in 1983, are the most common devices used in interventional cardiology today. Charles Thomas Stent (1807–1885), an English dentist who lent his name to a tooth mould. Charles Dotter used the word "stent" in 1963 to name endoluminal scaffolding devices. However, these interventions have created a new problem, restenosis.

Restenosis is a wound healing process occurring directly at the angioplasty balloon or stent site. It is believed that three processes cause restenosis: elastic recoil, neointimal hyperplasia, and negative vascular remodeling. Elastic recoil, or vessel spasm, occurs in the healthy (plaque-free) portion of the vessel within minutes after balloon expansion (angioplasty balloon or stent-expanding balloon). Elastic recoil causes a luminal cross-sectional area reduction of $\sim 50\%$, but only for a short time after the procedure. The second component of restenosis is neointimal hyperplasia resulting in new tissue growth occupying the microcracks and rapture within the plaque mass, in some patients this process can be overcompensating, filling more tissue within the vessel lumen thus compromising blood flow. The blood vessel wall (any vessel larger than capillaries) has three major layers called tunica, the innermost layer is the intima followed by a middle concentric layer called the media where mostly the smooth muscle cells reside, and then the outer most layer called the adventitia. The adventitia contains a connective tissue with mainly collagen and myofibroblasts. Some controversy still remains as to which cells are responsible for neointimal hyperplasia, media smooth muscle cells, or myofibroblasts migrating from the adventitia. In IVB, the prescription dose should reach the tunica adventitia to insure full therapeutic benefit. The third component of restenosis is negative remodeling. The term negative remodeling refers to contraction of the arterial wall following an arterial injury inflicted by an interventional procedure occurring slowly over the first 3–9 months after the angioplasty. Negative remodeling is believed to play a major factor in restenosis after a PTCA intervention. However, the stent is a mechanical scaffolding device that prevents negative remodeling. Hence, in-stent restenosis appears to derive almost exclusively from neointimal hyperplasia, even more than seen in balloon angioplasty. Schwartz and Holmes (4) provide a detailed discussion on restenosis and remodeling. Hall et al. (5) present on the radiobiological response of vascular tissue to IVB. (5).

Epidemiology and Clinical Trials

Restenosis is a very likely event (within few months of the initial intervention) and has been a frustrating problem in interventional cardiology. For example, > 1 million people worldwide had percutaneous coronary interventions in 2001 and of these $> 85\%$ received a stent. Restenosis occurred in $> 50\%$ of the stented patients ($\sim 425{,}000$ patients worldwide), with the United States share of $\sim 150{,}000$ patients.

Restenosis as measured using quantitative coronary angiography (QCA) is arbitrarily defined as a narrowing

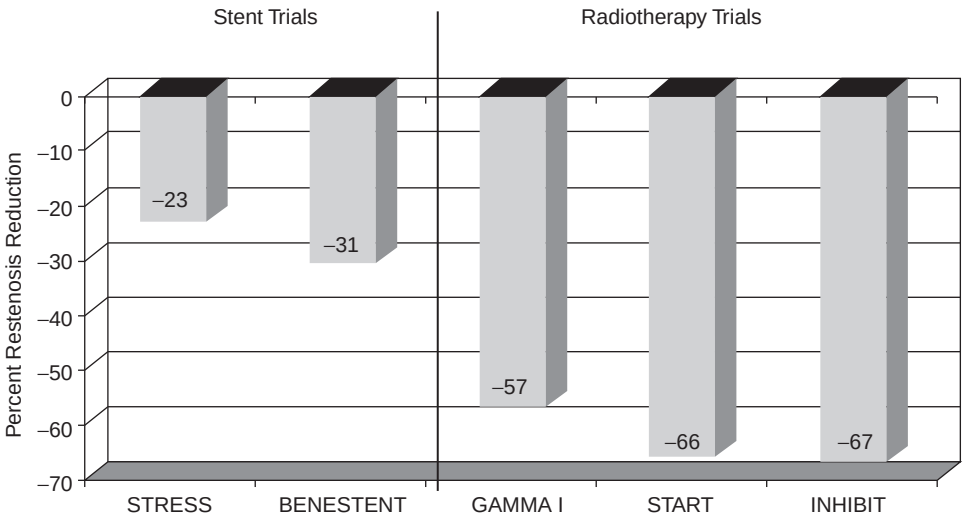


Figure 1. Percent restenosis reduction reported in bare-metal stent (STRESS and BENESTENT), and in intravascular brachytherapy clinical trials Gamma I (^{192}Ir), START (^{90}Sr/^{90}Y), and INHIBIT (^{32}P).

of the vessel lumen of at least 50% relative to the adjacent healthy vessel lumen ratio of minimum lumen diameter (MLD) to the reference lumen diameter (RLD). Mehran et al. (6) studied several risk factors from angiographic patterns of in-stent restenosis (ISR). The MLD, lesion length, and diabetes were found to be important factors predicting in-stent restenosis risk. Risk increased as a function of lesion length (10–40 mm) and decreased as a function of the MLD (2.5–4 mm diameter). Further, the ISR pattern is important where diffused restenosis has a larger risk those with a focal pattern.

The main complications of PTCA are acute vessel occlusion and late restenosis. Two important trials started in 1991, the North American STRESS (STent REStenosis Study) (7) and the European BENESTENT (BElgium Netherlands STENT trial) (8) using metal-bare stents (Palmaz-Schatz stent) transformed the practice of interventional cardiology. (Bare-metal stent is a term used here to make a distinction from the drug-coated stent briefly discussed in this article.) However, in-stent restenosis incidence of $>50\%$ was still observed. The introduction of IVB in the 1990s revamped the hope to eradicate restenosis. Condado et al. (9) in 1997 conducted the first small human trial showing reduction in restenosis over a 5 year follow-up period; this study, however, had some dosimetric issues. Definite multicenter double-blinded randomized clinical trials for IVB for the indication of in-stent restenosis are the GAMMA I (10), the START (Strontium-90 Treatment of Angiographic Restenosis Trial) (11), and the INHIBIT (INtimal Hyperplasia Inhibition with Beta Instent restenosis Trial) (12). With over 5000 patients treated with IVB, this modality has recently proven its safety and efficacy. As shown in Fig. 1, IVB (GAMMA I, START, INHIBIT) clinical trials had about twofold reduction of in-stent restenosis than found from the STRESS and BENESTENT bare-metal stent versus. PTCA trials. Initial problems in the IVB trials, like edge failure due to geographic miss and increased incidence of later thrombosis, were quickly remedied by increasing the radioactive source length and prolonged use of antiplatlet therapy. Table 1 is a detailed summary of these IVB clinical trials. Many other clinical IVB clinical trials were conducted using various isotopes, delivery system, and other clinical indications.

THEORY AND DETAILED DESCRIPTION OF IVB DEVICES

Based on the clinical trials discussed above, the Food and Drug Administration (FDA) granted a premarket approval (PMA) to three IVB devices. The Checkmate system (Cordis Corporation, a Johnson & Johnson Company, Miami Lakes, FL) using ^{192}Ir and BetaCath system using ^{90}Sr (Novoste Corp. Norcross, GA) were both approved on November 3, 2000. About 1 year later, the GALILEO Intravascular Radiotherapy System (Guidant Corp., Santa Clara, CA) using ^{32}P was also approved by the FDA. All of these devices are classified as catheter-based radiation delivery systems using sealed radioactive sources. Catheter-based means the source in the form of a source wire or a train of seeds (ribbon) is placed inside a closed-tip lumen catheter. The catheter is first placed into the target vessel and the source wire or ribbon is delivered or afterloaded using manual or computer-based delivery systems. The radiation safety considerations for these devices have been greatly discussed in the literature (13,14).

Other irradiation techniques were also investigated, including inflation of dilatation balloon catheter with radioactive liquid or gas; insertion of miniature X-ray tubes; implantation of radioactive stents; and postangioplasty external beam irradiation. These techniques did not make it to the market due to variable reasons including suboptimal efficacy, safety, or practicality. Table 2 summarizes several other isotopes and delivery systems investigated for IVB applications. Tables 3 and 4 list a few radiation characteristics for important gamma and beta sources for IVB.

Cordis Checkmate System

Checkmate is indicated by the FDA for the delivery of therapeutic doses of gamma radiation for the purpose of reducing in-stent restenosis. The system is for use in the treatment of native coronary arteries (2.75–4.0 mm in

Table 1. Summary of Pivotal IVB Clinical Trial Used to Obtain FDA Approval for In-Stent Restenosis in Native Arteries Indication

					Angiographic Restenosis		TLR	
Trial	Target Lesion	Source	Dose, Gy	Patients, *n*	Rad., %	Placebo %	Rad. %	Placebo, %
Gamma-1	In-stent (< 45 mm)	^{192}Ir ribbon	8–30	252	22	50 (6 months)	24	42 (9 months)
START	In-stent (< 20 mm)	^{90}Sr/^{90}Y Seed train	16–20 Gy at 2 mm	476	14	41 (8 months)	16	24 (8 months)
INHIBIT	In-stent (< 45 mm)	^{32}P wire	20 Gy at 1 mm into vessel wall	332	16	48 (9 months)	11	29 (9 months)

diameter and lesions up to and including 45 mm in length) with in-stent restenosis following percutaneous revascularization using current interventional techniques. Outside of the FDA approved indication, Waksman et al. (15) also examined the effects of intravascular gamma radiation in patients with in-stent restenosis of saphenous-vein bypass grafts and found favorable results.

Radioactive Source Ribbon. The Checkmate catheter-based brachytherapy system uses ^{192}Ir seeds that are preassembled in 6, 10, and 14 seed strand inside nylon ribbons (Best Medical International, Springfield, VA). Treatment lengths are 23, 39, and 55 mm for the 6-, 10-, and 14-seed ribbons, respectively (16). Iridium-192 has an average energy of 370 keV and a half-life of 73.83 days. The ^{192}Ir radioactive metal (30% Ir, 70% Pt) is 3 mm long and 0.1 mm in diameter encapsulated within a 3 mm long × 0.5 mm diameter stainless steel capsule. The seeds are placed inside a nylon ribbon with an interseed spacing of 1 mm. The overall ribbon length is 230 cm and the outer diameter is 0.76 mm (2.4 F). [1 French (F) = 1/π mm.] At both distal and proximal edges of the seed strand, radiopaque markers are placed for visualization under X rays. A nonradioactive dummy ribbon is preloaded inside the delivery catheter to provide reinforcement during shipping and to improve maneuverability during initial positioning of the catheter across the target lesion. The dummy ribbon has the same length and configuration of the radioactive source ribbon to aid in IVB therapy planning during the procedure. A source lumen plug is used to prevent the movement of the dummy ribbon inside the Checkmate delivery catheter during initial catheter placement via the femoral artery. Dosimetry characterization of the Checkmate source ribbon are discussed in the literature (17).

Table 2. Gamma and Beta Sources with Different Delivery Systems Investigated for Intravascular Brachytherapy Applications

Gamma Delivery Systems	Beta Delivery Systems
^{192}Irseed train	^{32}P - wire, stent, balloon
^{125}I-stent	^{90}Sr/^{90}Y - seed train
^{103}Pd-stent, wire	^{90}Y - wire
^{131}Cs-stent	^{188}W/188Re - wire, balloon-liquid
^{99m}Tc-liposome–liquid	^{186}Re - balloon-liquid
	^{133}Xe - balloon-gas
	^{48}V - stent (positron)[a]
	^{62}Cu - balloon-liquid
	^{106}Ru/^{106}Rh - wire
	^{144}Ce/^{144}Pr - wire
	^{68}Ge/^{68}Ga balloon-liquid (positron)

[a] A positron is an electron with a positive charge.

Table 3. Average Energy and Half-Life of Important Gamma Emitters Investigated for Intravascular Brachytherapy Applications

Isotope	Ave Energy, keV	$T_{1/2}$, day
^{192}Ir	370	73.83
^{125}I	28	59.4
^{103}Pd	21	16.97
^{131}Cs	30	9.69

Delivery Catheter. This is a single lumen catheter with a distal rapid exchange tip and a closed-ended source lumen for isolation from patient blood contact. Both the radioactive source and the nonradioactive dummy ribbon use this lumen to reach the target. A single radiopaque marker at the distal end of the source lumen is to aid in catheter placement under fluoroscopy. A guidewire is used to guide the catheter along the tortuous pathway into the coronary artery. The guidewire exits the catheter 4 mm from the distal tip of the catheter. The overall length of the Checkmate catheter is 230 cm with a usable length of 145 cm. At the distal portion of the catheter, the outer diameter is 3.7 F (0.049 in.), which is deliverable with 7 F (mm) or larger guiding catheter.

Table 4. Maximum Energy and Half-Life of Important Beta Emitters Investigated for Intravascular Brachytherapy Applications

Isotope	Max Energy, keV	$T_{1/2}$
^{32}P	1710	14.3 day
^{90}Sr/^{90}Y	2280	29.1 year
^{90}Y	2280	64 h
^{188}W/^{188}Re	2120	69.4 day
^{186}Re	1090	90.6 h
^{133}Xe	360	5.3 day
^{48}V	696	16 day
^{62}Cu	2930	9.74 min
^{106}Ru/^{106}Rh	3540	371.6 day
^{144}Ce/^{144}Pr	3000	284.9 day

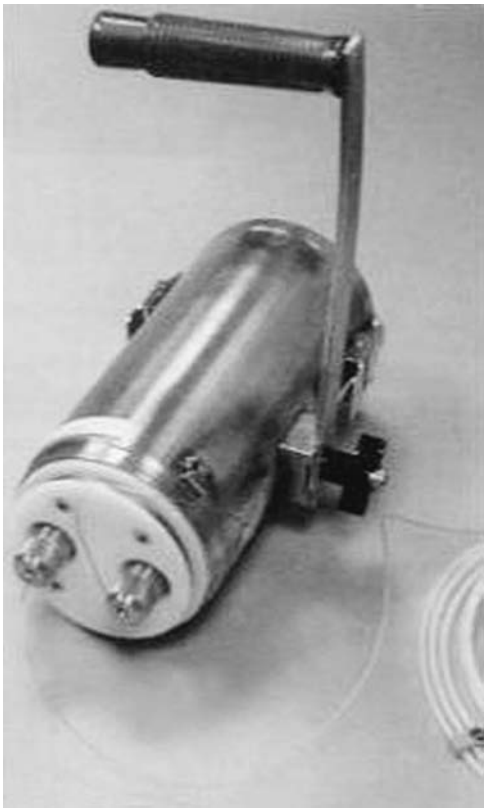


Figure 2. Checkmate delivery device (Cordis Corporation, a Johnson & Johnson Company, Miami Lakes, FL) is mainly a lead shielded cylinder housing both the radioactive ^{192}Ir source ribbon and dummy ribbon.

Delivery Device. As shown in Fig. 2, the Checkmate delivery device is mainly a lead shielded cylinder housing both the radioactive ^{192}Ir source ribbon and the dummy ribbon. This is a simple device where the proximal end of the source ribbon (nonradioactive part) protrudes from the proximal end of the delivery device and is coiled when not in use. When in use, a threaded cap on the distal end of the delivery device is replaced with a luer connector, which is connected to the hub of the delivery catheter. The source ribbon is pushed forward by hand from the proximal end of the delivery device into the delivery catheter.

Dosimetry. The Checkmate IVB system dosimetry was initially based on intravascular ultrasound (IVUS). From these images, the distance from the center of the IVUS catheter to the outer edge of the media tissue, called the external elastic membrane (EEM) is measured. A minimum of three axial images is taken along the stented vessel segment to determine the maximum and minimum distance from the source to the EEM. The dwell time is then calculated to insure that 8 Gy is delivered to the EEM farthest from the source, provided that no >30 Gy is delivered to the closest EEM. The IVUS-based dosimetry was later simplified to prescribe a fixed dose of 14 Gy at a distance of 2 mm from the centerline of the source. This provided a logistic solution to shorten procedure time and to spread the use of this system since many of catheterization labs in the United States do not have IVUS image modality.

Novoste BetaCath

The first generation BetaCath is a 5.0 F (1.59 mm) system. This system was indicated by the FDA to deliver beta radiation to the site of successful percutaneous coronary intervention for the treatment of in-stent restenosis in native coronary arteries with discrete lesions of $\leq$ 20 mm in length using the 30 or 40 mm system and for longer lesions up to 40 mm using a longer source train (60 mm) in a reference vessel diameter ranging from 2.7 to 4.0 mm. The second generation BetaCath is a 3.5 F (1.17 mm) system, which has an equivalent radioactivity to the 5 F (1.59 mm) system, but is smaller in diameter and fits easily inside a (1.91 mm) guide catheter. This system is intended to deliver beta radiation.

The 3.5 F (1.17 mm) BetaCath system has three main components, the ^{90}Sr source train, the β-Rail 3.5 F delivery catheter, and the 3.5 F delivery device.

Radioactive Source. Strontium-90/Yttrium-90 is a pure beta emitter with any energy spectrum with maximum energy of 2.27 MeV and an average of 0.934 MeV. The long half-life (29.1 years) simplifies treatment planning due to the almost unchanged dose rate of the device during the life cycle of the device (6 month). Each seed is 2.5 mm long and 0.38 mm in diameter for the 3.5 F system (0.64 mm seed diameter for the 5 °F system), manufactured by AEA Technology GmbH, Germany or BEBIG Isotopen- und Medizintechnik GmbH, Berlin, Germany. The radioactive source train consists of a wire jacketed "train" of 12 (30 mm source train), 16 (40 mm source train), or 24 (60 mm source train). The jacketed design was a major improvement over the initial design to eliminate seed movement thus providing uniform dose distribution. A radiopaque mark is placed one each side of each source train to provide visualization under fluoroscopy. Dosimetry characterization of the BetaCath sources are discussed in the literature (18,19).

Delivery Catheter. Only details of the second-generation delivery catheter, β-Rail 3.5 F (1.17 mm), will be discussed. This is a closed-end catheter with a total length of 180 cm (Fig. 3). A longer catheter called β-Rail 3.5 F XL delivery catheter has an overall length of 267 cm if desired. The catheter has a guidewire exit port at 1 cm from the distal tip, that is, a rapid exchange design. This catheter accommodates all source train lengths (30, 40, or 60 mm) that reach a most distal radiopaque marker located inside the delivery catheter. The β-Rail delivery catheter is preloaded with an Indicator of Source Train (IST), this Novoste terminology is used for the nonradioactive dummy source to aid in the measurement and positioning of the delivery catheter to insure adequate radiation coverage. The IST includes two radiopaque markings to delineate 30, 40, and 40 mm source lengths. At the proximal end of the delivery catheter, a proprietary connector is provided to insure a secure connection to the delivery device described next.

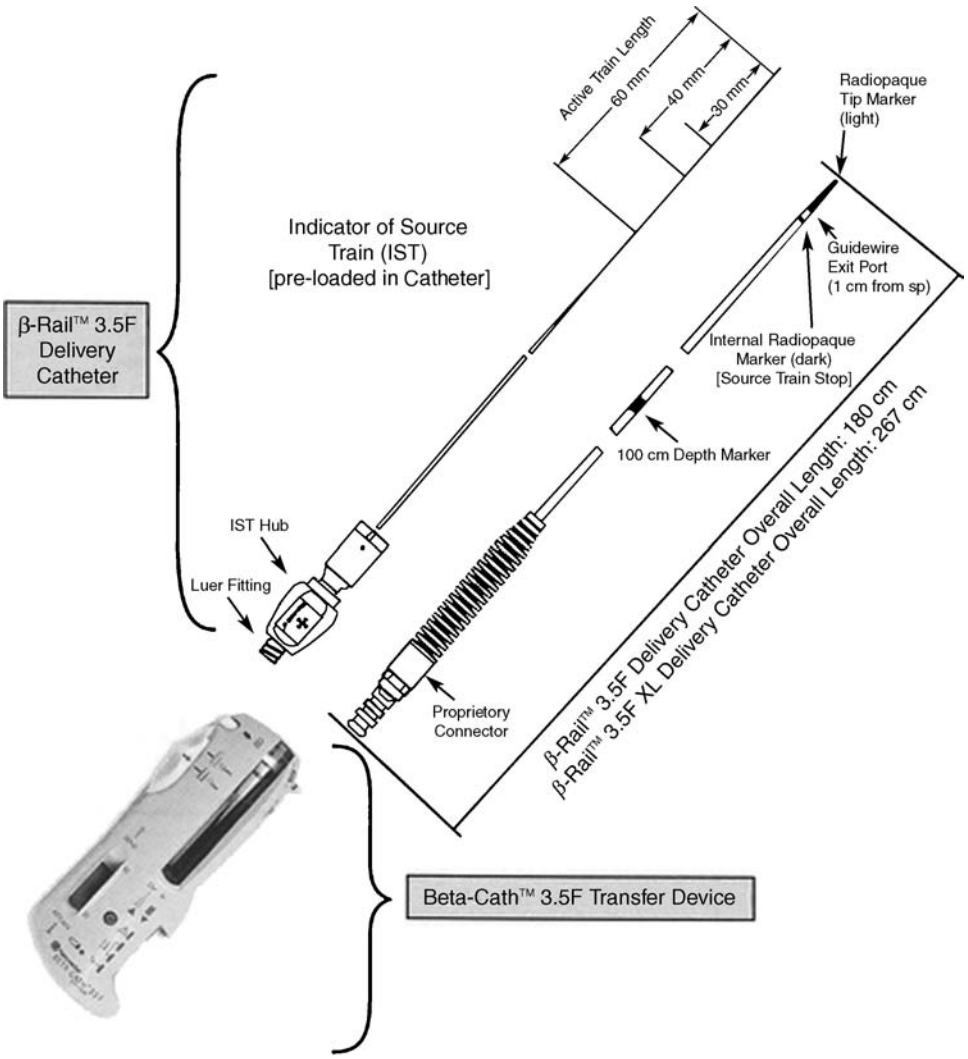


Figure 3. The β-Rail 3.5 F delivery catheter (Novoste, Norcross, GA) is a closed-end catheter with a total length of 180 cm. A longer catheter called β-Rail 3.5 F XL delivery catheter has an overall length of 267 cm is available. (Courtesy of Novoste Corporation.)

Delivery Device. This is a handheld battery-powered device used to store the radioactive source train (only one length). Hence, three separate delivery devices are needed to accommodate the 30, 40, and 60 mm source trains described above. The source train is sent into and returned from the delivery catheter using hydraulic pressure. To insure safe attachment of the delivery catheter, a connector lock latch is provided at the exit port of the delivery device. Several electronic pressure sensors are used to provide the operator with feedback on the pressure required using a saline-filled syringe to send, hold, or return the source train. Two source train position indicator lights (Green: In/Amber: Out), adjacent to the source chamber-viewing window (see Fig. 4). A fluid control lever controls the fluid flow and direction of the source train movement. A treatment counter tracks the number of procedures or test runs, a maximum of 125 transfers is allowed.

Dosimetry. The BetaCath IVB system dose prescription is the simplest out of the three systems discussed. The dose prescription is given relative to the source axis at 2 mm radial distance. The recommended dose is 18.4 Gy for a measured reference vessel diameter < 3.35 mm, but > 2.7 mm; and 23 Gy for a diameter > 3.35 mm, but < 4.0 mm. Vessel diameters < 2.75 mm or > 4.0 mm can be treated with this system; however, this is considered an off-label use of the device as defined by the FDA. The appropriate source train length (30, 40, or 60 mm) is selected after measuring the injured length using angiography and adding a margin on the distal and proximal side of a minimum of 5 mm.

Guidant Galileo

This IVB system is the only computer-based device. The GALILEO Intravascular Radiotherapy System (Guidant Corp., Santa Clara, CA) consists of three main components: the GALILEO ^{32}P Source Wire, the GALILEO Centering Catheter, and GALILEO Source Delivery Unit (SDU). The first generation product used a 27 mm long ^{32}P source and spiral centering catheter. The second generation, called GALILEO III uses a 20 mm long ^{32}P source, a trichannel centering catheter, and an automated high precision stepping algorithm.

The first generation GALILEO was indicated to deliver beta radiation to the site of successful percutaneous coronary intervention for the treatment of in-stent restenosis

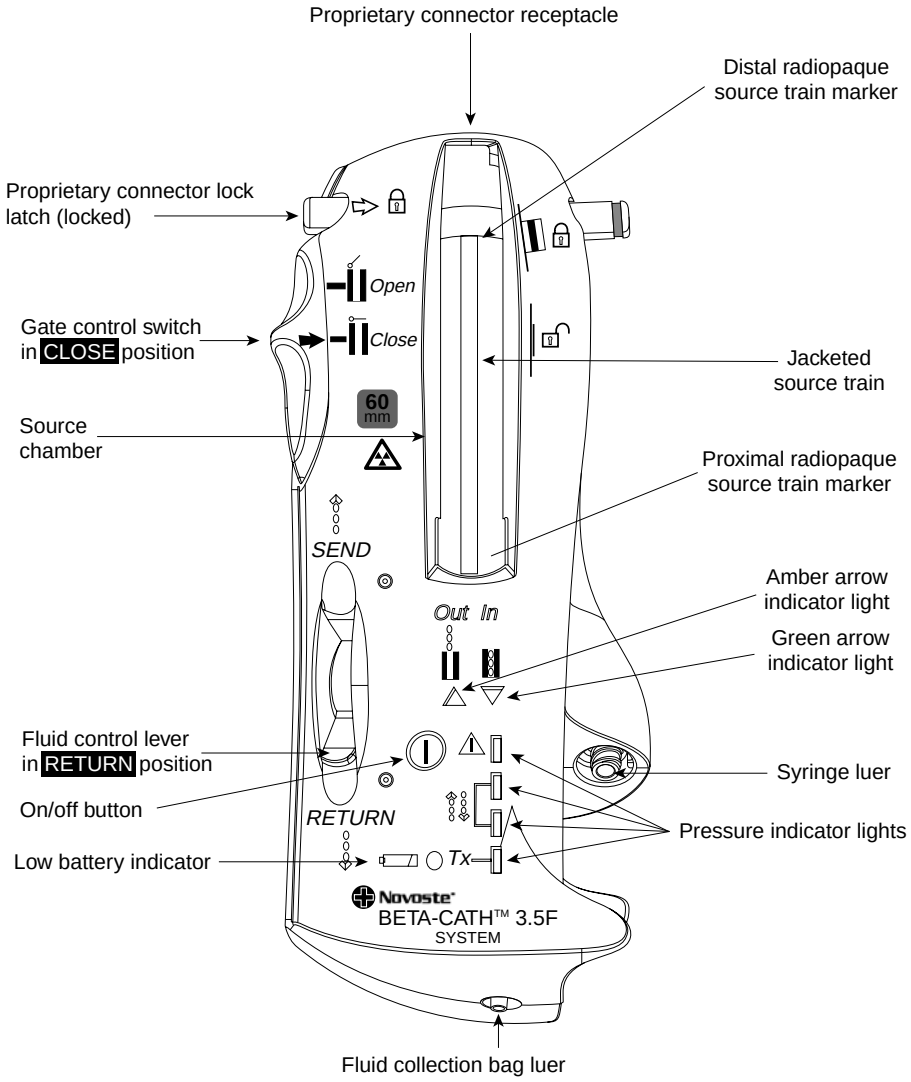


Figure 4. The BetaCath 3.5 F system (Novoste, Norcross, GA) is a handheld battery-powered device used to store the radioactive source train (only one length of 30, 40, or 60 mm source trains). Source train transfer depends on hydraulic pressure. Several electronic pressure sensors are used to provide the operator with feedback on the pressure required using a saline-filled syringe to send, hold, or return the source train. (Courtesy of Novoste Corporation.)

in native coronary arteries with discrete lesions ≤ 47 mm in reference vessel diameter 2.4–3.7 mm. The second generation GALILEO III system extended the indication to treat injured arterial length up to 52 mm.

Radioactive Source. The active wire contains linear solid-form phosphorus 32 (^{32}P) in ceramic glass fiber sealed in the distal end of a flexible nitinol (NiTi) hypotube, which is welded to a nitinol wire (total wire length is 2430 mm). The nominal active length is 27 mm (first generation system) 20 mm (second generation system). Both source wires have an outer diameter of 0.46 mm. Phosphorus-32 is a pure beta-emitting isotope with a maximum energy of 1.71 MeV, an average energy of 0.690 MeV, and a half-life of 14.28 days. The active wire can be used in multiple procedures for ~ 4 weeks (two half-lives). The active wire has two 1 mm tungsten X-ray markers, one proximal and one distal from the source, for visualization (see Fig. 5). A dummy source is used before the active wire is delivered to verify positioning, to check for kinks and catheter obstructions that could prevent the active wire from reaching the treatment site, and to achieve accurate positioning at the treatment site. Similar to the active wire, the dummy source also has two tungsten markers, making it visually identical to that of the active wire. Dosimetry characterization of the Galileo sources are discussed in the literature (20,21).

Delivery Catheter. This is a dual-lumen catheter with a spiral-shaped (first generation) or triloped balloon (second generation) to provide source centering within the lumen to improve dose homogeneity (see Fig. 6a and b). Such balloon profiles are designed to center the source within the lumen and to allow distal and side-branch perfusion during the dwell time that can take up to 10 min. This would make the procedure more tolerable for patients. The design of the second generation was found to provide better perfusion than the spiral design, in particular for the longer balloons to treat longer lesions. One lumen allows the automatic advancement of the source wire. At the proximal end of this lumen, a key connector attaches the delivery catheter to the delivery device. The key connector has a special code that reads using an optical sensor at the entry port to automatically determine the balloon length and the number of dwell positions based on the centering catheter used. The distal end of this lumen is closed to prevent contact of the source wire with the patient blood. The second lumen

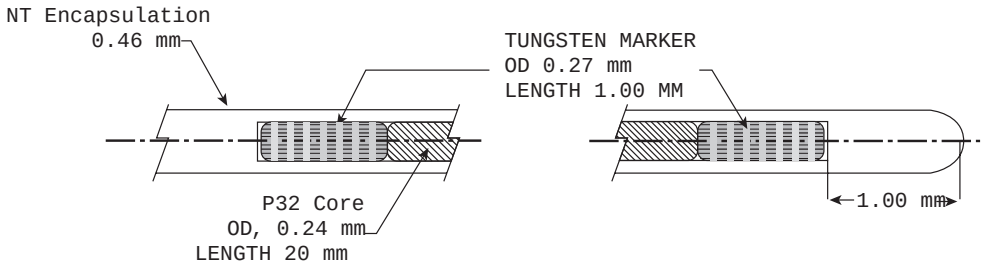


Figure 5. Cross-section of the ^{32}P source wire (Guidant Corporation, Santa Clara, CA) is shown. The radioactive core has a nominal length of 20 mm and 1 mm tungsten X-ray markers, one on each side of the core. All dimensions are in millimeters. (Courtesy of Guidant Corporation: GALIELO system is no longer manufactured or for sale.)

allows inflation (recommended pressure is 4 atm) and deflation of the centering balloon. This lumen terminates in a luer-lock connector allowing the attachment of standard inflation devices. A third lumen is 5 mm long at the most distal tip of the catheter; this is used to place the delivery catheter over a standard 0.014 in. (0.36 mm) coronary guide wire using a Rapid Exchange approach. Radiopaque markers located at the distal and proximal end of the balloon allows proper placement of the delivery catheter under fluoroscopy. Proximal shaft markers are located at 95 and 105 cm to aid in gauging catheter position relative to the tip of a brachial or femoral guiding catheter, respectively. The triloped GALILEO III centering catheter is provided with a balloon diameter of 2.5, 3.0, and 3.5 mm and balloon lengths of 32 and 52 mm. (Balloon length is defined as the distance between radiopaque balloon markers and does not include balloon tapers that extend beyond these markers). The MLD determines the appropriate centering catheter balloon diameter to use in the artery segment being treated (see Table 5). The lesion length determines the appropriate centering catheter length to use (see Table 6).

Delivery Device. The delivery device or the SDU has three main components: the head, base, and cartridge.

The front of the SDU head (Fig. 7) includes the touch screen monitor, status indicator lights, and housing for the cartridge. It also contains the manual retract wheel, the cartridge key port, the catheter key port and catheter eject button, and the red STOP button. On the back of the SDU head are the touch screen tilt lever, the swivel handle, and the system key port. The SDU head houses two motor drives—a primary motor and a battery-operated emergency retract motor. The emergency-retract motor works automatically if the primary motor fails to retract the active wire.

The SDU base provides a stable foundation for the SDU head and allows the unit to be transported easily.

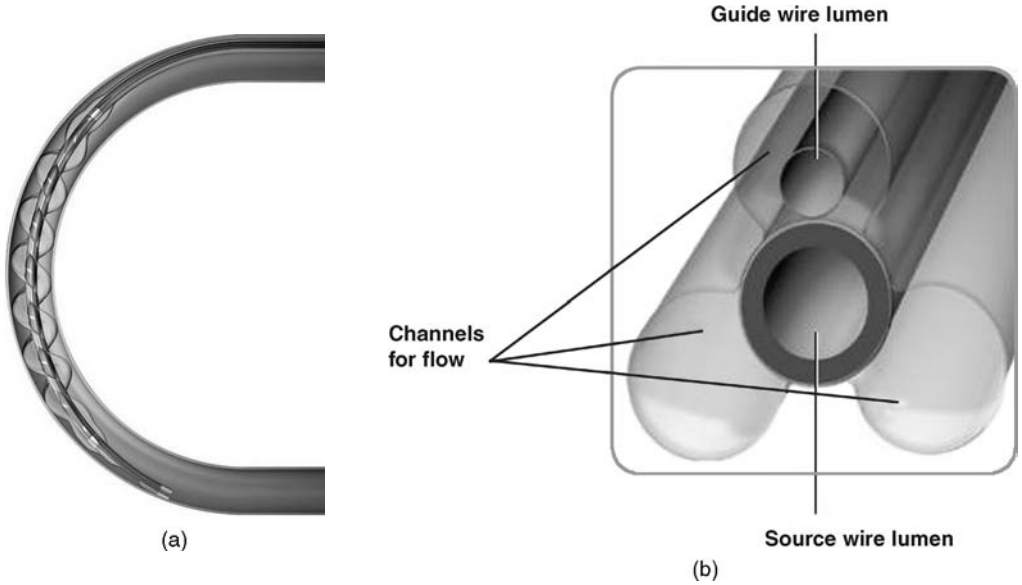


Figure 6. (a) Galileo first generation spiral centering catheter (Guidant Corporation, Santa Clara, CA) is shown inside a 3 mm diameter curved lumen, balloon is inflated to 4 atm with saline to provide optimal source centering and distal blood perfusion (Courtesy of Guidant Corporation—GALIELO system is no longer manufactured or for sale.) (b) Cross-section of a Galileo second generation trilobed centering catheter (Guidant Corporation, Santa Clara, CA) is shown. Source lumen is central and the three lobes are inflated to 4 atm with saline to provide optimal source centering and distal blood perfusion. Note guide wire lumen inside one of the lobes. (Courtesy of Guidant Corporation—GALIELO system is no longer manufactured or for sale.)

Table 5. Balloon Diameter Selection for the Guidant Galileo Centering Catheter as a Function of the Measured MLD

Balloon Diameter, mm	MLD, mm
2.5	2.25–2.75
3.0	2.75–3.25
3.5	3.25–3.7

Table 6. Balloon and Equivalent Source Length Selection

Balloon Length, mm	Injured Arterial Length, mm	Equivalent Source Length, mm
32	≤ 32	40
52	33–52	60

Components of the base include the head release, the handle bar, the emergency compartment, the wheels and wheel locks, and the power cord port. The emergency compartment contains the equipment necessary to handle an emergency, including the emergency safe, the emergency wire cutter, and the emergency tongs.

The cartridge, which is inserted into the SDU head, contains the active (^{32}P source) wire, the dummy wire, and the operating software. It also contains the catheter key port, the tungsten safe, which shields the active wire when not in use, and the wire drive mechanisms. Figure 8 is an example of the GALILEO software screen of the countdown clock.

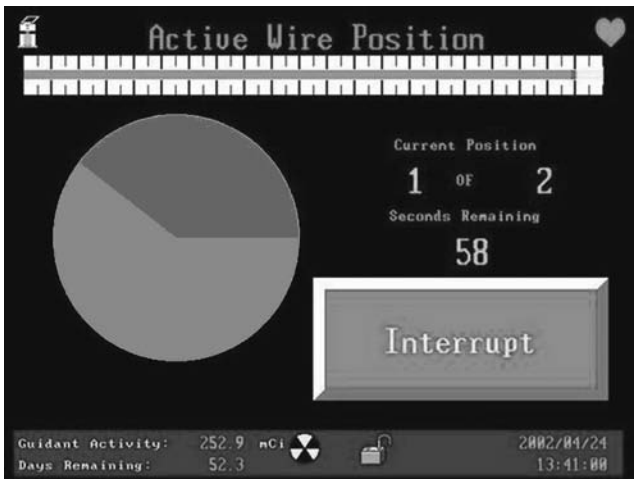


Figure 8. A screen shot of the Galileo software of the countdown clock for a treatment with two-source positions using the automatic source positioning system. (Courtesy of Guidant Corporation—GALIELO system is no longer manufactured or for sale.)

Dosimetry. Based on the measured minimal lumen diameter and lesion length via fluoroscopy, online QCA, or IVUS, a proper centering balloon size is chosen. Also the lumen diameter of the nondiseased vessel is measured immediately proximal and immediately distal to the treatment area. The average of these two diameters is the RLD. The GALILEO prescription point for radiation delivery is 1 mm beyond the RLD. The SDU automatically calculates the dwell time required to deliver the prescribed dose of radiation (20 Gy) at the prescription point. The GALILEO

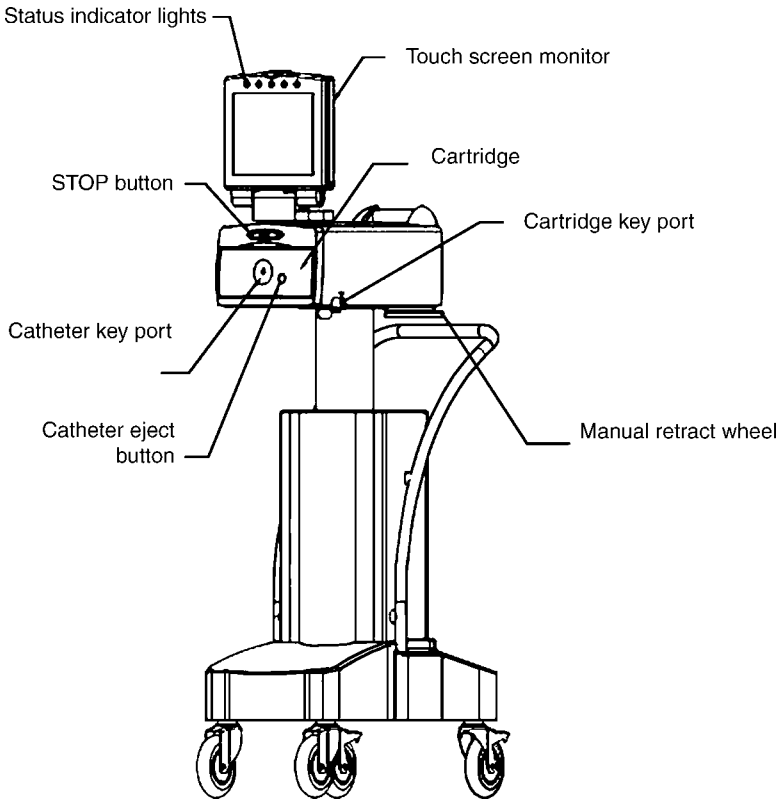


Figure 7. A front view of the Galileo source delivery system. It includes the Touch Screen Monitor, Status Indicator Lights, and housing for the Cartridge. It also contains the Manual Retract Wheel, the Cartridge Key Port, the Catheter Key Port and Catheter Eject Button, and the red STOP Button. (Courtesy of Guidant Corporation–GALIELO system is no longer manufactured or for sale.)

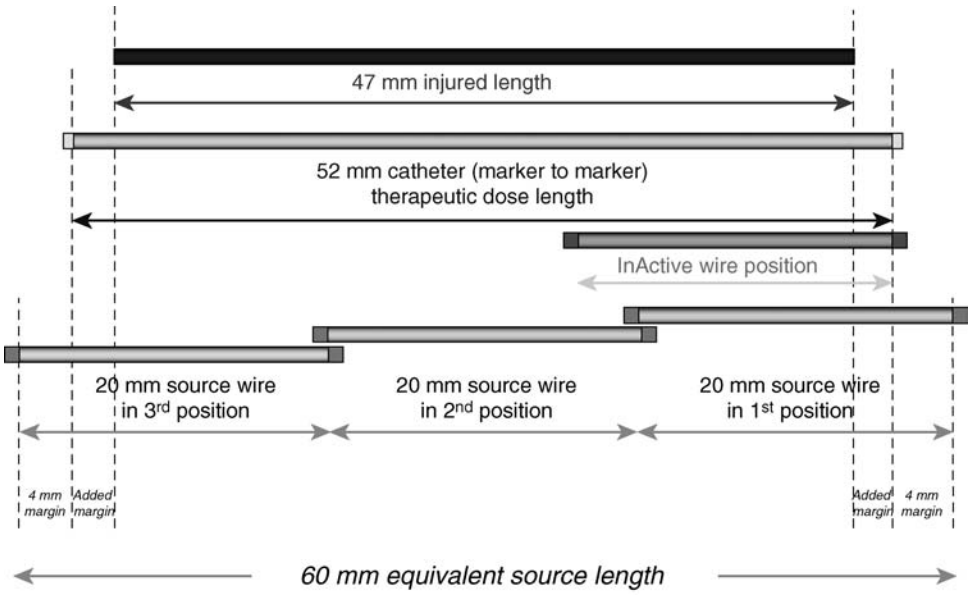


Figure 9. An example of 60 mm equivalent source length resulting from stepping twice a 20 mm ^{32}P source to treat a 47 mm lesion. Note the added margin of 2.5 mm to the minimal margin of 4 mm to provide adequate radiation coverage.

SDU automatically steps the 20 mm ^{32}P source from most distal position to the most proximal position to yield an equivalent source length sufficient to cover the injured length with a minimal margin of 4 mm (Table 6). Equivalent source length (ESL) is defined as the total source length that will be result by stepping the 20 mm active source in tandem, either in two dwell positions (40 mm equivalent source length) or in three dwell positions (60 mm equivalent source length). Figure 9 is an example of 60 mm ESL to treat a 47 mm lesion, note the added margin of 2.5 mm to the minimal margin of 4mm to provide adequate radiation coverage.

DESIGN CONSIDERATION FOR IVB DEVICES AND DOSIMETRY

For beta emitters (^{90}Sr/^{90}Y, ^{32}P) used currently in intravascular brachytherapy, the prescribed dose is greatly influenced by several perturbation factors. These perturbation factors are divided into two categories; the first is *applicator* dependent and the second is *patient anatomy* dependent. Important applicator perturbation factors include the placement of a guidewire during irradiation, the source lumen eccentricity within the applicator and centering capability with the vessel lumen, use of contrast, X-ray markers, and source stepping precision (if injured lengths longer than the source radioactive length are treated). Patient perturbation factors include vessel anatomy (size, curvature, and cross-section eccentricity), plaque morphology (composition, density, and spatial distribution), and stent type. High energy gamma emitters like) ^{192}Ir (J&J Checkmate system) are influenced by such factors as well, but with almost negligible magnitudes (22), hence the next discussion will focus on the does distributions perturbation of the beta sources only.

Applicator Dependent Perturbations

The applicator dependent factors discussed in this report are specific for two commercially available IVB beta source systems, that is, Novoste (^{90}Sr/^{90}Y) and Guidant (^{32}P) systems. Five factors are discussed and considered most important, but are not inclusive, (*1*) guidewire, (*2*) source lumen eccentricity within the applicator and centering capability within the vessel lumen, (*3*) use of contrast, (*4*) X-ray markers, and (*5*) source stepping precision (manual vs. automatic).

Guide Wire Perturbation. The guide wire (GW) is used to navigate through the cardiovascular arteries during common interventional procedures, such as balloon angioplasty, stenting, and IVB. Commonly used guide wires are made of stainless steel and have a solid cross-section with diameter of 0.014 in. (0.36 mm). Several authors have reported on the dose perturbation due to the guide wire in IVB (23–25). Figure 6b depicts the cross-section of 3.0 mm GIII centring catheter with a 0.014 in. (0.36 mm) guide wire inside the upper lope. The distance from the center of the GW and source axis is not fixed in this design and the GW location can vary (shown at closest distance from the source lumen center). Also Fig. 10 depicts the location of the guide wire inside the 5 Fr BetaCath catheter. Shih et al. (24) provides a detailed study on dose perturbation as a function of the GW position relative to the source axis. A dose perturbation

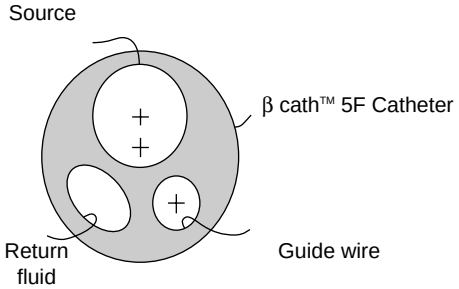


Figure 10. Cross-section of the BetaCath 5 F delivery catheter. Note the location of the guide wore lumen relative to the source lumen. (Courtesy of Novoste Corporation.)

factor (DPF), defined as the ratio of the doses with and without the presence of a guidewire was introduced to quantify the effects.[100] The authors reported a DPF of up to 70% behind a GW for the beta sources. The dose reduction for the beta sources was found to be dependent on the guidewire location. For example, the dose reduction was 10% higher for a stainless steel guidewire located at 0.5 mm than that for the guidewire at 2 mm from the central axis of the source. The portion of the target volume affected (shadowed) dosimetrically by the guidewire was reduced when the guidewire was positioned farther away from the source. The shadow volume (in which the dose reduction occurs) can be reduced by up to 45% as the guidewire is moved away from the source axis from 0.5 to 2 mm.

Fluhs et al. (25) measured the GW [0.014 in. (0.3556 mm), stainless steel] perturbation using a plastic scintillator. The authors pointed that the insertion of a GW into the catheter close to the beta source causes a large angular asymmetry of the radiation emission. For a guide wire positioned eccentrically to the catheter the dose reduction is dominantly limited to a region of some 20° around the angle defined by catheter centerline and guide wire. At the catheter surface the maximal dose reduction in this region was found to be ($30 \pm 2\%$, DPF = 10%). At the typical dose prescription depth of 2 mm from the source axis the shielding effect decreased to ($24 \pm 2\%$) (or DPF = 76%). This value is remarkably larger than the dose reduction caused by any typical stent design.

Source Lumen Eccentricity. Sehgal et al. (26) reports on the dosimetric consequences of source centering (eccentricity) within the arterial lumen as one potentially important factor for the uniform delivery of dose to the arterial tissue. In this study, they have examined the effect of source centering on the resulting dose to the arterial wall from clinical intravascular brachytherapy sources containing ^{32}P and ^{90}Sr/^{90}Y. Monte Carlo simulations using the MCNP code (described in Advanced Topics section below) were performed for these catheter-based sources with offsets of 0.5 and 1 mm from the center of the arterial lumen in homogenous water medium as well as in the presence of residual plaque. Three different positions were modeled and the resulting dose values were analyzed to assess their impact on the resulting dose distribution. The results are shown in Table 7. The debate on the importance of centering of beta emitters used in IVB to treat native coronary vessels has been extensively published (27).

Contrast Perturbation. Contrast agents with high atomic number materials are usually injected into blood vessels to help in the determination of lesion location and to verify source placement during the IVB procedure (small fraction of the entire dwell time). Common contrast agents like Omnipaque and Hypaque are discussed. Omnipaque contains ~25% of iodine (in mass), and Hypaque contains ~23% of iodine. Iodine has an atomic number (Z) of 53. Nath et al. (22) discussed the perturbation factor of these contrast agents on ^{32}P and ^{90}Y IVB sources when the contrast is injected directly in the blood stream, however, this paper did not provide the average DPF over the treatment dwell time. Mourtada used a Monte Carlo simulation of the Galileo III centering catheter (Fig. 6b) to calculate the average dose reduction at 1 mm tissue depth if the three catheter lobes are filled with saline (as recommended by the product instruction for use) and 50:50 Omnipaque contrast. The results are reported in Table 8. The simulations were done for the three different sizes of the GALILEO III centering balloon.

Table 7. Results Are Reported at a Radial Distance of 2 mm from the Coronary Artery Lumen Center[a]

Offset from Center, mm	^{32}P, %	^{90}Sr/^{90}Y, %
0.5	−40 to +70	−30 to +50
1.0	−65 to +185	−50 to +140

[a]Data is from Ref. 26.

Table 8. Average DPF at 1 mm into the Vessel Wall when the Galileo III Centering Catheter is Filled with 50:50 Omnipaque Contrast Agent[a]

Balloon diameter, mm	DPF, %
2.5	−2.9
3.0	−4.8
3.5	−7.6

[a]The contrast remains in the balloon for the entire ^{32}P irradiation dwell time. Data calculated by Mourtada using MCNP Monte Carlo code (unpublished data).

X-Ray Markers Perturbation. The GALILEO III centering catheter has distal and proximal X-ray markers made of 90% Pt and 10% Ir (effective density is 21.6 g cm^{-3}). The X-ray markers are 0.635 mm long and the inner and outer diameter are 0.394 and 0.432 mm, respectively. The GALILEO 20 mm source first position inside the GIII centering catheter is positioned 4 mm beyond the proximal edge of the distal X-ray marker to provide adequate margins. Figure 11a depicts the 20 mm source wire (red) and distal X-ray marker (gold). Using the Monte Carlo simulation MCNP, the dose distribution in water around a 20 mm ^{32}P source was calculated with and without the distal X-ray marker. Figure 11b depicts the two-dimensional (2D) isodose map with the distal X-ray marker in place. As expected due to scattering, the X-ray marker perturbation is reduced quickly as a function of depth. For a 3.0 mm diameter vessel example, the intimal surface maximum DPF is 23% and at the prescription depth of 1 mm into the vessel wall, the maximum DPF is 15%. The effect of the radiopaque marker in the BetaCath ^{90}Sr90/^{90}Y system has not been reported, but it is expected to be minimal since the catheter markers are along the side of the radioactive seed train.

Source Stepping. Lesions that are longer than available IVB sources require stepping the source (i.e., tandem positioned) from mostly distal to most proximal lesion segment to provide adequate radiation coverage. First-generation clinical beta-source systems have gained FDA approval for clinical indication for focal lesions (< 22 mm) (12,28). For injured lesions > 22 mm, but ≤ 47 mm, the manual tandem positioning (MTP) technique was investi-

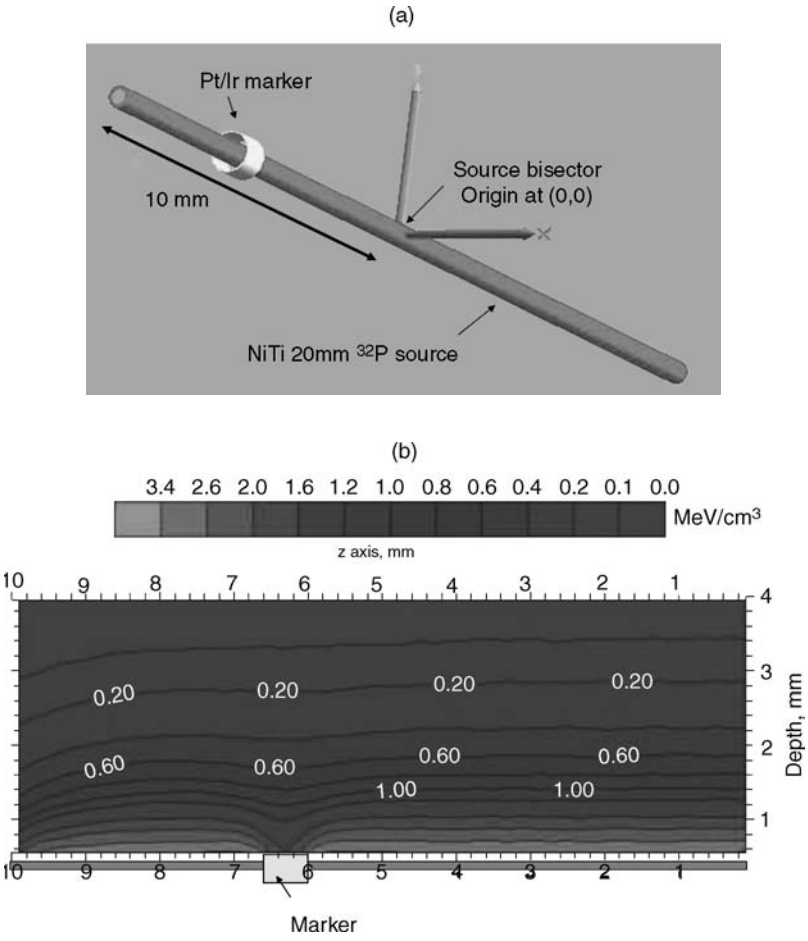


Figure 11. (a) Schematic of the distal X-ray marker of the Galileo second generation centering catheter (GIII) relative to nominal position of the 20 mm ^{32}P source inside the GIII catheter. (b) Energy deposited per unit volume in water (unit: MeV cm^{-3}) calculated using MNCNPX Monte Carlo code for the geometry shown in Fig. 11a. The origin of the coordinate system is located at the bisector of the 20 mm ^{32}P source. The proximal edge of the distal X-ray marker (gold) is located at 6 mm from the source origin. Note the large dose perturbation at closer depth (< 1 mm).

gated in the INHIBIT clinical trial using the 27 mm ^{32}P source (12). From the INHIBIT data analysis; for the 56 patients treated who had a tandem-positioning procedure (and the core lab had reported the size of gap or overlap), 44% had no gap or overlap. But, 19 and 11% of the patients had a 1 (32% increase in dose at junction) and 2 mm (56% increase in dose at junction) overlap respectively. Only one patient (1.9%) had a 1 mm gap (32% decrease in dose at junction). Hence, a 2 mm overlap and 1 mm gap are defined as the upper limits allowed for tandem positioning. Crocker et al. (29) also investigated the MTP procedure with 30 and 40 mm ^{90}Sr/^{90}Y source trains, and concluded from their data that the MTP technique was safe from both a dosimetric and a clinical point of view. However, Coen et al. (30) published a retrospective evaluation of the accuracy of manual multisegmental irradiation with 30 and 40 mm ^{90}Sr/^{90}Y source trains for irradiation of long (re)stenotic lesions in coronary arteries, following PTCA. They concluded that the positioning inaccuracy of MTP caused unacceptable dose inhomogeneities at the junction between source positions, and the procedure was not recommended. Coen et al. (30) suggested using longer line sources or source trains, or preferably an automated stepping source to insure reliable and safer technique for treatment of long lesions. Table 9 lists the dose perturbation at stepping junction due to an overlap or gap at the reference depth of 2 mm in water for GALILEO ^{32}P and Novoste ^{90}Sr/^{90}Y sources.

To reduce MTP dosimetric errors and to allow adequate coverage of radiation to longer injured lengths using the same source, an afterloader can be used to automatically step the radiation source to yield a longer equivalent source length. The only IVB system capable of this is the second generation GALILEO automated stepping system using a 20 mm ^{32}P source wire (33).

Patient Anatomy Dependent Perturbations

Patient perturbation factors discussed in this article include (*1*) vessel geometry (size, curvature, tapering,

Table 9. Dose Perturbation at Junction Due to an Overlap or Gap at the Reference Depth of 2 mm in Water for GALILEO ^{32}P and Novoste ^{90}Sr/^{90}Y Sources

Size of Overlap or Gap, mm	^{32}P[a], %	^{90}Sr/^{90}Y[b], %
0	0	0
0.5	±17	
1	±32	±23
2	±56	±44
3	±75	±60
5	±91	±80

[a] Guidant Corporation Data to Ref. 31.
[b] See Ref. 32.

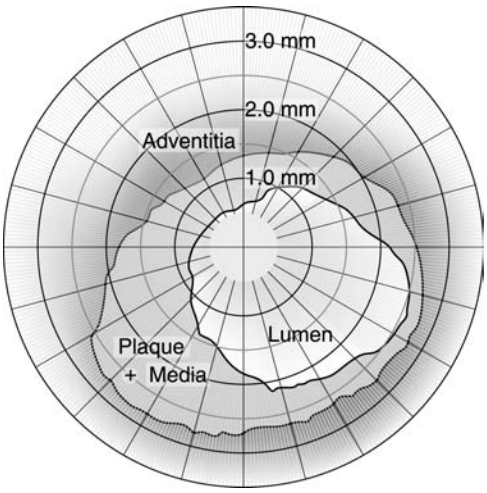


Figure 12. A typical vessel cross-section is eccentric, and the vessel center does not necessarily lie on the lumen center.

and cross-sectional eccentricity); (*2*) plaque morphology (composition, density, and spatial distribution); (*3*) stenting.

Vessel Geometry Perturbation. The native coronary vessel diameter has a range from 2 to 4 mm. Due to plaque formation and tapering, the lumen diameter is a variable. It is expected that the dose perturbation factor to be worse for larger vessels, especially for the BetaCath IVB system whose dose does not have active centering. Also, a typical vessel cross-section is eccentric, and the vessel center does not necessarily lie on the lumen center (Fig. 12). As pointed by Kaluza et al. (27) in reality, plaque thickness can vary from 0.1 to 2.3 mm. The average eccentricity index was 6.38 ± 5.95 in the 59 PREVENT patients. The intimal hyperplasia of in-stent restenosis complicates the vessel cross-section. Mehran et al. (6) developed an angiographic classification of in-stent restenosis mainly under two categories: focal and diffused (6).

Another important vessel geometry factor is curvature. Xu et al. (34) studied the effect of curvature on ^{32}P beta dosimetry. As expected, the curvature causes an increase in dose in the inner surface (concave side) of the coronary vessel and a decrease in dose in the outer surface (convex side). For a maximum theoretical bend of 180°, the dose increases by as much as 20% along the inner radial distance, but decreased by as much as 20% along the outer radial distance compared to the dose along a straight wire. The authors concluded that for curvatures normally encountered in a clinical situation, the dose rate was changed by $< 5\%$.

Plaque Morphology Perturbation. The artery mostly consists of normal healthy tissue, but may also contain plaque, whose density may be unknown. Plaque is a material that develops inside the artery over time and is considered responsible for blockage of the artery. Plaque may range widely in histologic structure, density, and chemical composition. Density is expected to depend on the plaque's collagenous matrix and degree of calcification. Rahdert et al. (35) measured the density and calcium concentration

Table 10. DPF Values at 2 mm Radial Distance for ^{32}P and ^{90}Sr/^{90}Y Sources[a]

Plaque Density, g cm^{-3}	DPF, ^{32}P	DPF, ^{90}Sr/^{90}Y
No plaque	1.0	1.0
1.45	0.93	0.97
1.55	0.91	0.96
3.10	0.70	0.83

[a]The plaque layer has a thick-ness of 0.2 mm for all the different cases. The relative error is within 5% for the given dose rate values. (See Ref. 26).

in 13 cadaveric plaque specimens. This study concluded that based on the plaque calcification, the density range is between 1.25 and 1.5 $g \cdot cm^{-3}$.

Several studies on dose perturbation due to plaque have been reported in the literature for catheter-based beta sources (36–38). Nath et al. (36) assumed a 1 mm thick plaque with cortical bone density of 1.84 $g \cdot cm^{-3}$ and 27% Ca composition. Both values are relatively higher than those reported by Rahdert et al. (35). For this extreme condition, however, Nath et al. (36) reported ~ 0.8 mm reduction in penetration for ^{90}Sr/^{90}Y and ^{32}P beta sources when the calcified plaque was located next to the source, and by ~ 0.9 mm when the plaque was located 1 mm away from the source.

Li et al. reported $\sim 30\%$ reduction due to plaque for the Novoste ^{90}Sr/^{90}Y source (37). The modeled calcified plaque density range in this study was 1.2–1.60 $g \cdot cm^{-3}$ with a nominal density of 1.45 $g \cdot cm^{-3}$ as reported for B100 bone equivalent by ICRU Report 26 (39). Li et al. (37) calculated the DPF as a function of the radial distance from the source axis. Sehgal et al. (26) reported on the perturbation due to concentric plaque with constant thickness of 0.2 mm and three different densities as shown in Table 10 for both the GALILEO ^{32}P and Novoste ^{90}Sr/^{90}Y sources. Comparing with Li et al. (37) the 0.2 mm thick plaque with 1.45 $g \cdot cm^{-3}$ DPF results are similar.

Stent Perturbation. Even though a stent is not part of the actual vessel anatomy, it is assumed that an expanded stent is predislodged into the intima and surrounded by new tissue growth as a result of in-stent neointimal hyperplasia. Hence, the stent becomes part of the diseased vessel segment. In a few instances, a vessel could receive a second or even a third stent, as part of the intervention of a recurrent in-stent restenosis.

For all stent types, a similar behavior can be observed. At a distance of 0.5 mm and more the typical overall dose reduction does not exceed a value of 5–15% (40). In the close vicinity of the stent struts, an increased dose reduction effect reaches up to some 30–40%. The large dose reduction directly behind the stent struts are caused by the absorption effect that is not compensated by scattering contributions from regions outside of the shielded area until a depth of ~ 0.5 mm from the strut surface (41).

Recently, a new stent made of cobalt chromium L-605 alloy (CoCr, $\rho = 9.22$ $g \cdot cm^3$) (MULTI-LINK VISION) was introduced as an alternative to the commonly used 316L stainless steel stent design (SS, $\rho = 7.87$ $g \cdot cm^{-3}$) (MULTI-LINK PENTA). Mourtada and Horton (42) used the Monte

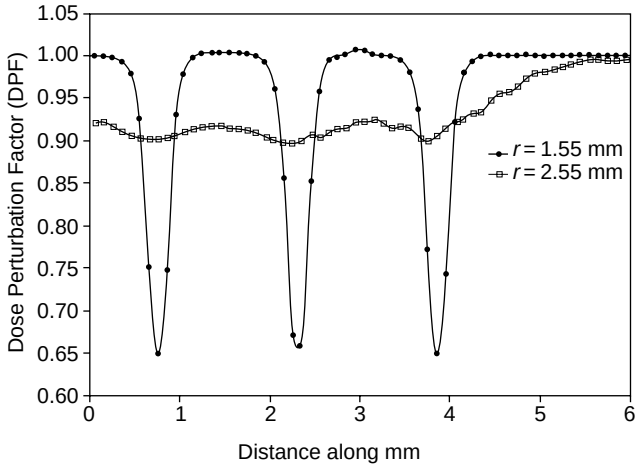


Figure 13. A 3 mm diameter stainless steel stent DPF along the source axis ($x = 0$ is the coordinate system origin at source bisector). The parameter $r = 1.55$ mm is the radial distance from the source axis and is the centroid of the scoring bin directly behind the stent (score bin radial thickness is 0.1 mm). The parameter $r = 2.55$ mm is the centroid of the scoring bin at the 2.5 mm prescription point. Data calculated using MCNPX Monte Carlo code. (Reference 42 with permission from Medical Physics journal.)

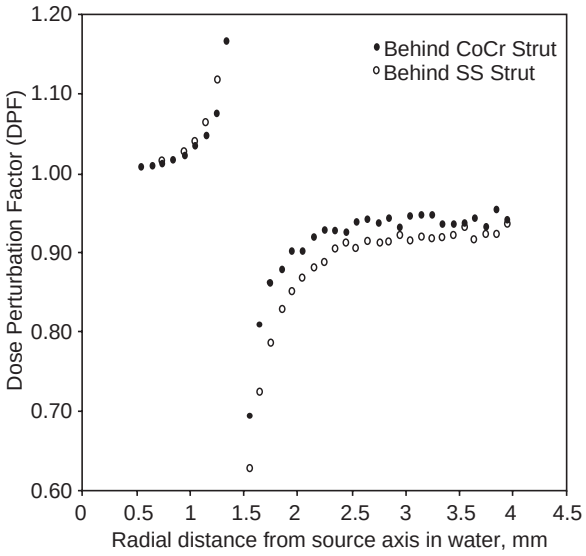


Figure 14. Dose perturbation factor as a function of the radial distance from source axis in water and through a strut located 1.5 mm from the source axis (expanded inside a 3.0 mm vessel model). (Reference 42 with permission from Medical Physics journal.)

Carlo code MCNPX to compare the dose distribution for the ^{32}P GALILEO source in CoCr and SS 8 mm stent models. The DPF, defined as the ratio of the dose in water with the presence of a stent to the dose without a stent, was used to compare results. Both stent designs were virtually expanded to diameters of 2.0, 3.0, and 4.0 mm using finite element models (ABQUS Inc., Pawtucket, RI). The complicated strut shapes of both the CoCr and SS stents were simplified using circular rings with an effective width to yield a metal/tissue ratio identical to that of the actual stents. The mean DPF at a 1 mm tissue depth, over the entire stented length of 8 mm, was 0.935 for the CoCr stent and 0.911 for the SS stent. The mean DPF at the intima (0.05 mm radial distance from the strut outer surface), over the entire stented length of 8 mm, was 0.950 for CoCr, and 0.926 for SS. The maximum DPFs directly behind the CoCr and SS struts were 0.689 and 0.644, respectively. Figures 13 and 14 depict the dose profiles behind the stainless steel stent as an example. The authors concluded that although the CoCr stent has a higher effective atomic number and greater density than the SS stent, the DPFs for the two stents are similar because the metal/tissue ratio and strut thickness of the CoCr stent are lower than those of the SS stent.

ADVANCED TOPICS

Figure 15 is the general dosimetry characterization paradigm of brachytherapy sources. This requires three important steps including the experimental measurement of the dose rate ($cGy^{-1} \cdot s$) as a function of distance from the source using a calibrated dosimeter, a measurement of the source contained activity (SI unit is the Becquerel = Bq = 1 decay s^{-1}). The normalized measure dose rate to the measured contained activity can then be compared to a theoretical calculation such as a Monte Carlo simulation, which is inherently has dose rate units per particle emitted, that is, contained activity. This section will discuss briefly an example of each component that is important in the IVB dosimetry paradigm.

Theoretical Dosimetry: Monte Carlo Simulations

The Monte Carlo technique used for IVB dosimetry is considered the most accurate theoretical tool, particularly suited in handling the complex interactions of the emitted beta particles with the surrounding medium. Fox has published a review of IVB (43), including a rather thorough review of the theoretical dosimetry applied to these sources. Detailed reviews and discussion of the Monte Carlo method can also be found in the literature (44–46). Monte Carlo codes utilized for IVB dosimetry include CYLTRAN from the Integrated Tiger Series (ITS version 3.0, Sandia National Laboratory, Albuquerque, NM), the MCNP series (MCNP4C and MCNPX, Los Alamos

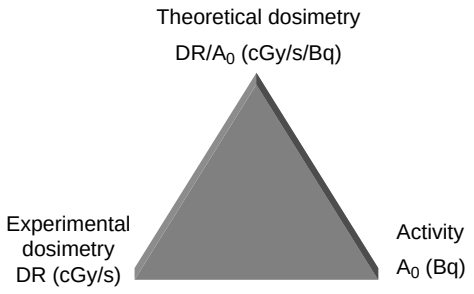


Figure 15. A general dosimetry characterization paradigm of brachytherapy sources.

National Laboratory, Los Alamos, NM), the EGS series (EGS4 Stanford Linear Accelerator Center, Stanford, CA, and EGSnrc National Research Council of Canada, Ottawa, Ontario, Canada), and PENELOPE (University of Barcelona, Barcelona, Catalonia, Spain).

Theoretical modeling is being increasingly used to supplement measurements in IVB. The Monte Carlo method, for example, is based on the idea that if all materials and dimensions of a problem, and all the probabilities of the various possible radiation interactions are known, then emitted particles can be tracked and scored as they are transported from the source through the media of the problem. A random number generator is used to select emitting source element location, emitted particle energy and direction, and results of various interactions and secondary radiations as the particle is tracked to either absorption or out of the problem boundaries. Obviously, the more "histories" (emitted particles) are tracked, the more accurate is the resulting calculation. With the advent of faster and faster computers, Monte Carlo calculations are becoming more and more attractive for determining dose distributions for brachytherapy sources in complex geometries. Often, the combination of a Monte Carlo calculation, which yields dose rate per unit contained activity, and a contained activity measurement, will give better accuracy than a dosimetric measurement. An excellent review of the Monte Carlo method has recently been published (44).

MC simulations of electron transport, for example, beta-particle transport, are usually different from those used in photon or neutron transport, for which the simulated radiation history is followed individually based on conventional methods since uncharged radiation interactions are characterized by relatively infrequent isolated collisions. For example, in photon transport, the distance to the next photon interaction is sampled from the attenuation coefficient distribution; and the change in attenuation coefficient as the photon crosses material boundaries is modeled. The type of interaction is sampled from the appropriate relative probabilities. The history of each photon is continued from collision to collision until the photon either is absorbed, escapes the problem boundary, or its energy falls below a chosen cutoff threshold at which the remaining energy of the photon is locally deposited.

For high energy electrons, such a detailed history is not practical for energies > 100 keV, because many individual elastic and inelastic Coulomb collisions per history are generated through the media resulting in very long computational time. Instead, a "condensed history" is used, where the electron trajectories are divided into many path segments (47). For each path segment, the net angular deflection and the net energy loss are sampled from relevant multiple-scattering distributions. The choice of the step size is important for accuracy and is chosen with conflicting requirements. On the one hand, the steps should be short enough that (*1*) most of the electron history steps are completely inside the boundary of a predefined surface, so that the use of multiple-scattering theories of unbounded media is valid; (*2*) the energy loss is, on average, small within a step; and (*3*) the net angular deflection is, on average, small so that the path within the step is approximated by a straight line. On the other hand, the step size should be large enough to contain a sufficient number of collisions per step to justify the use of the multiple-scattering theories and to limit the number of steps per history to reduce computing time. Further discussion can be found in the literature (45).

Experimental Dosimetry: Radiochromic Film Measurements

Both MD-55 and HD-810 radiochromic dye films (RCF) (GAFChromic type, Nuclear Associates, Carle Place, NY) are widely used in IVB dose-field measurement due to their superior spatial resolution. Also, RCF is used instead of other types of films mainly for its linear dose response. A full description of both film types is reported by AAPM Task Group 55 (48) For beta field measurements, it is recommended that the RCF dosimeter be calibrated using the same $^{90}Sr/^{90}Y$ ophthalmic applicator (New England Nuclear S/N 0258) calibrated at the National Institute of Standards and Technology (NIST), Gaithersburg, Maryland (20). Polystyrene or other tissue-equivalent materials are used to fabricate high precision blocks for the film measurements. Each block has a hole with a diameter slightly larger than the source diameter to reduce positional error (in IVB 0.1 mm could translate to 13% error in the measured dose rate). Several blocks are made with nominal depths (distance from center of hole to block surface) ranging from 0.5 to 5 mm. Actual depths must be verified using a traveling microscope or an optical comparator. At each of these depths, several radiochromic films should be exposed for a range of times to gain a good image of the radiation field. Digitization of film is typically done with a high resolution 2 scanning densitometer (Pharmacia LKB) using a 633 nm laser (HeNe) with a 100 μm diameter spot size and a 40 μm minimum step size. Alternatively, scanning is done using a high resolution (242×375) CCD densitometer (CCD100, Photoelectron, Lexington, MA) with a 665 nm LED array and a 160×200 μm pixel size. To account for optical-density growth as a function of time after exposure, film readout should be done 72 h postirradiation for both the calibration and experimental films. The net optical density measurements of each film were converted into a 2D dose map as shown in Fig. 16. Estimated dose uncertainties for radiochromic film are ± 15.6 % ($k = 2$); the individual components of this uncertainty are shown in Table 11.

Other radiation dosimeters mostly lack the spatial resolution of submillimeters required in IVB. However, recently Amin et al. (49) proposed using a polyacrylamide gel (PAG) dosimeter and a high field 4.7 T MRI scanner for IVB. The get/scanner final in-plane resolution of 0.4×0.2 mm is approaching the film resolution, but not quite. The authors confirmed that both absorbed dose and dose distributions for high gradient vascular brachytherapy sources can be measured using PAG, but the disadvantages of gel manufacture and the need for access to a high resolution scanner suggests that the use of radiochromic film is the method of choice (49).

Determination of Contained Activity. The activity content of this wire must be known in order to relate the measurements and calculations of the absorbed-dose spatial

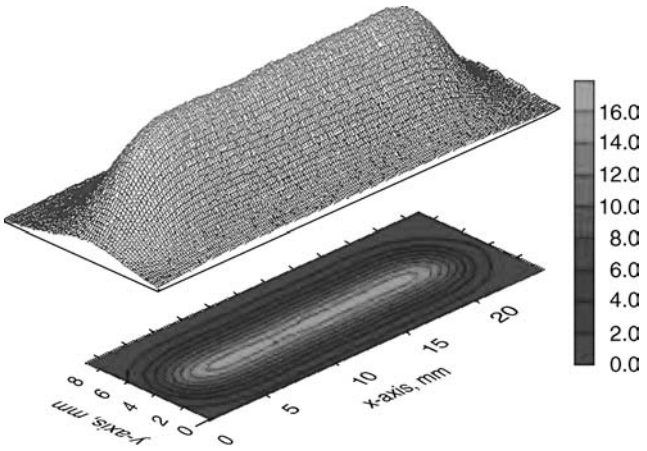


Figure 16. A 2D HD810 radiochromic film image dose (Gy) distribution for a ^{32}P source, measured in a plane parallel to the source's longitudinal axis at 1.97 mm radial distance from the source axis in the polystyrene block.

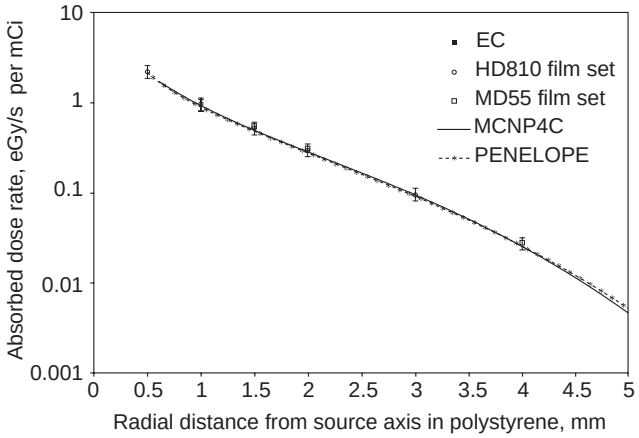


Figure 17. Measured data (Film and EC: extrapolation chamber) depth dose curve plotted along with Monte Carlo estimates from MCNP4C and PENELOPE. The error bars estimate the 95% confidence interval. All data are measured or calculated in polystyrene. (Reference 21 with permission from Medical Physics journal.)

distribution. For the 27 mm ^{32}P source design, Mourtada et al. briefly described the determination of the absolute contained activity from the original work of Collé (50). The contained activity was then related to a calibration factor for the NIST Capintec. The CRC-12 ionization chamber (51). Similar work was done on the Novoste ^{90}S/^{90}Y seed and other beta sources used for IVB to establish radioactivity standards by NIST, Gaithersburg, MA (52,53).

For example, Fig. 17 is the Galileo 20 mm ^{32}P source wire measured radiochromic film (MD55 and HD810) depth dose curve plotted along with Monte Carlo estimates from MCNP4C and PENELOPE. The error bars estimate the 95% confidence interval. All data are measured or calculated in polystyrene (21).

Beyond IVB Treatments of Heart Disease, Other Applications, and Future Roles

In the United States, there are $\sim$8–12 million patients affected with peripheral vascular disease. An estimated 600,000 interventional procedures are performed each year, including percutaneous transluminal angioplasty (PTA), bypass surgery, and amputation. Percutaneous transluminal angioplasty restenosis rates are high with a success rate of $<$ 23% at 6 months follow-up. Intravascular brachytherapy has been investigated to reduce restenosis in the superficial femoral and popliteal arteries after PTA. The main IVB clinical trial for peripheral vessel is the Peripheral Artery Radiation Investigation Study (PARIS) using the Nucletron Ir-192 HDR source (mHDR v2). The PARIS study used the Nucletron micro-Selectron high dose rate (HDR) afterloader and the Guidant PARIS centering catheter (10–20 cm long and 4–8 mm diameter) (54). More recently, the MOBILE clinical trial a ^{90}Sr/^{90}Y source and the Corona gas-filled centering is being investigated (Novoste Corporation) (55).

Other possible applications of IVB include treatment of recurrent narrowing of the arteriovenous (AV) dialysis graft (56), renal artery stenosis, transjugular intrahepatic porto-systemic (TIPS) stenosis, carotid artery (57), and subclavian vein stenosis. Further application of IVB might be for treatment of atrial fibrillation, a most common cardiac arrhythmia. It is proposed that IVB radiation dose can electrically isolate ectopic foci located mostly in the adventitia of the pulmonary vein (PV), which are responsible for atrial fibrillation episodes (58). The IVB approach might alleviate undesirable side effects (PV stenosis due to heating) of rf ablation; a commonly used treatment modality to ablate myocardial tissue.

Intravascular Brachytherapy in the Drug-Eluting Stent Era

The recent introduction of drug-eluting stents (DES) in the interventional cardiology arena has a tremendous impact on the IVB practice. By incorporating antiproliferative agents onto the surface of the stent, neointimal hyperplasia

Table 11. Estimated Relative Uncertainties of the Radiochromic-Film Dose Interpretations per Measured Unit Activity

Uncertainty Component	Relative Standard Uncertainty, %
Calibration of the NIST standard ^{90}Sr/^{90}Y calibration source	6
Response of the film exposed to the calibration source	3
Response of the films exposed to the source under test	3
Activity calibration	2.6
Combined standard uncertainty	7.8
Expanded uncertainty ($k = 2$)	15.6

occurring within the stent is markedly reduced. Stents coated with agents, like Sirolimus, Paclitaxel, Tacrolimus, Everolimus, and so on, when compared to bare-metal stents, had shown remarkable reduction in binary restenosis and target vessel revascularization (TVR) rates in several clinical trials (31,59).

As discussed earlier in this article, IVB has demonstrated its safety and efficacy in limiting recurrence of in-stent restenosis with positive long-term follow-up outcome. However, the utility of IVB is being rethought in relation to its use after the placement of drug eluting stents. It is expected that the role of IVB will decrease primarily due to the simplicity of placement of drug eluting stents and the relative complexity of performing intravascular brachytherapy. The published pivotal clinical restenosis rates from both the Paclitaxel (TAXUS) (32) and Sirolimus (RAVEL, SIRIUS) drug eluting stents suggests that the need to perform an IVB will be in < 1 in 20 patients (60–62). Recently reported registry data from Europe and the United States demonstrates similar efficacy of DES to that of IVB for treatment of in-stent restenosis (63,64). However, DES technology still has to go through further investigations in more complex lesions and higher risk patients in the general population to appreciate its full potential. Other agents with potential benefits like Statins, local gene therapy, and further innovations in polymer technology (biodegradable polymers, multiple-drug release polymers) are in under evaluation (31).

DEFINITION OF TERMS

MACE: Major Adverse Cardiac Events, a composite of death, MI, target, or repeat lesion revascularization.

Angiographic Binary Restenosis: Stenosis of 50% or more of the luminal diameter.

TLR and TVR: Target Lesion and Vessel Revascularization: Describes a rate measuring how many stented lesions had to be retreated, due to clinically driven restenosis, given a specific time period.

MLD: Minimal lumen diameter, the smallest diameter of an artery in a specified segment.

RLD: Reference lumen diameter, the average of two diameters, the lumen diameter of the nondiseased vessel immediately proximal and distal to the target treatment area.

Lumen Diameter: The inner diameter of an artery in a specified segment.

Neointimal Hyperplasia: Wound-healing response to arterial injury that leads to restenosis.

Late Loss: A cardiology term referring to the angiographic measurement of neointimal hyperplasia. It's one of the most important indicators of long-term efficacy in coronary intervention.

Percutaneous Transluminal Coronary Angioplasty (PTCA): A method of treating blood vessel disorders that involves the use of a balloon catheter to enlarge the blood vessel and thereby improve blood flow.

Restenosis: Narrowing of a vessel dilated by angioplasty or other interventional procedure.

In-stent Restenosis (ISR): Narrowing of a vessel after a stent is in place; this may be acute due to a thrombus formation or late (few months) due to wound healing process (neointimal hyperplasia and remodeling).

BIBLIOGRAPHY

Cited References

1. Gruentzig AR. Seven years of coronary angioplasty. Z Kardiol 1984;73 (Suppl.) 2:159–160.
2. Dotter CT, Buschmann RW, Mckinney MK, Rosch J. Transluminal expandable nitinol coil stent grafting: Preliminary report. Radiology 1983;147:259–260.
3. Cragg A, et al. Nonsurgical placement of arterial endoprostheses: A new technique using nitinol wire. Radiology 1983;147: 261–263.
4. Schwartz RS, Homes DR. Restenosis and remodeling. In: Waksman R. editor. Vascular brachytherapy. Armonk, (NY): Futura Publishing Company; 1999.
5. Hall EJ, Miller RC, Brenner DJ. Radiobiological principles in intravascular irradiation. Cardiovasc Radiat Med 1999;1:42–47.
6. Mehran R, et al. Angiographic patterns of in-stent restenosis: Classification and implications for long-term outcome. Circulation 1999;100:1872–1878.
7. Fischman DL, et al. A randomized comparison of coronary-stent placement and balloon angioplasty in the treatment of coronary artery disease. Stent restenosis study investigators. N Engl J Med 1994;331:496–501.
8. Serruys PW, et al. A comparison of balloon-expandable-stent implantation with balloon angioplasty in patients with coronary artery disease. Benestent study group. N Engl J Med 1994;331:489–495.
9. Condado JA, et al. Long-term angiographic and clinical outcome after percutaneous transluminal coronary angioplasty and intracoronary radiation therapy in humans. Circulation 1997;96:727–732.
10. Leon MB, et al. Localized intracoronary gamma-radiation therapy to inhibit the recurrence of restenosis after stenting. N Engl J Med 2001;344:250–256.
11. Popma JJ, et al. Randomized trial of $^{90}Sr/^{90}Y$ beta-radiation versus placebo control for treatment of in-stent restenosis. Circulation 2002;106:1090–1096.
12. Waksman R, et al. Use of localised intracoronary beta radiation in treatment of in-stent restenosis: The inhibit randomised controlled trial. Lancet 2002;359:551–557.
13. Balter S. A health physics perspective. In: WR, editor. Vascular brachytherapy. New York: Futura Publishing Company; 1999.
14. Barish R. Radiation safety considerations for intravascular brachytherapy. In: Balter S, Chan RC, Shope TB, editors. Intravascular brachytherapy and fluoroscopically guided interventions. Madison (WI): Medical Physics Publishing; 2002.
15. Waksman R. Intravascular gamma radiation for in-stent restenosis in saphenous-vein bypass grafts. N Engl J Med 2002;346:1194–1199.
16. Jani S, Massullo V, Tripuraneni P, Teristein P. The ^{192}Ir radioactive seed ribbon. In: Waksman R, Serruys P, editors. Handbook of vascular brachytherapy. London: Martin Dunitz Ltd; 2000.
17. Chiu-Tsao ST, et al. Verification of ^{192}Ir near source dosimetry using gafchromic film. Med Phys 2004;31:201–207.
18. Roa DE, et al. Dosimetric characteristics of the novoste beta-cath $^{90}Sr/Y$ source trains at submillimeter distances. Med Phys 2004;31:1269–1276.

19. Soares CG, Halpern DG, Wang C-K. Calibration and characterization of beta-particle sources for intravascular brachytherapy. Med Phys 1998;25:339–346.
20. Mourtada FA, Soares CG, Seltzer SM, Lott SH. Dosimetry characterization of ^{32}P catheter-based vascular brachytherapy source wire. Med Phys 2000;27:1770–1776.
21. Mourtada F, et al. Dosimetry characterization of a ^{32}P source wire used for intravascular brachytherapy with automated stepping. Med Phys 2003;30:959–971.
22. Nath R, Yue N, Weinberger J. Dose perturbations by high atomic number materials in intravascular brachytherapy. Cardiovasc Radiat Med 1999;1:144–153.
23. Li XA, Shih R. Dose effects of guide wires for catheter-based intravascular brachytherapy. Int J Radiat Oncol Biol Phys 2001;51:1103–1110.
24. Shih R, Hsu WL, Li XA. Dose effect of guidewire position in intravascular brachytherapy. Phys Med Biol 2002;47:1733–1740.
25. Fluhs D, et al. The influence of guiding equipment and stents on the beta dose distribution in the brachytherapy of in-stent restenosis. Cardiovasc Radiat Med 2001;2:241–245.
26. Sehgal V, Li Z, Palta JR, Bolch WE. Dosimetric effect of source centering and residual plaque for beta-emitting catheter based intravascular brachytherapy sources. Med Phys 2001;28:2162–2171.
27. Kaluza GL, et al. Targeting the adventitia with intracoronary beta-radiation: Comparison of two dose prescriptions and the role of centering coronary arteries. Int J Radiat Oncol Biol Phys 2002;52:184–191.
28. Suntharalingam M, et al. Clinical and angiographic outcomes after use of 90strontium/90yttrium beta radiation for the treatment of in-stent restenosis: Results from the stents and radiation therapy 40 (Start 40) registry. Int J Radiat Oncol Biol Phys 2002;52:1075–1082.
29. Crocker I, et al. Treatment of long, diffuse, in-stent restenotic lesions with beta radiation using strontium 90 and sequential positioning "pullback" technique: Procedural details and clinical outcomes. J Invasive Cardiol 2001;13:782–787.
30. Coen VL. Inaccuracy in manual multisegmental irradiation in coronary arteries. Radiother Oncol 2002;63:89–95.
31. Fattori R, Piva T. Drug-eluting stents in vascular intervention. Lancet 2003;361:247–249.
32. Stone GW, et al. One-year clinical results with the slow-release, polymer-based, paclitaxel-eluting taxus stent: The taxus-IV trial. Circulation 2004;109:1942–1947.
33. Waksman R, et al. Beta radiation delivered via an automatic stepping device to inhibit recurrence of diffuse in-stent restenosis: Clinical and angiographic results of the multicenter galileo inhibit clinical study. Circulation (Suppl II) 2001;104:II–509.
34. Xu Z, et al. The investigation of ^{32}P wire for catheter-based endovascular irradiation. Med Phys 1997;24:1788–1792.
35. Rahdert DA, et al. Measurement of density and calcium in human atherosclerotic plaque and implications for arterial brachytherapy. Cardiovasc Radiat Med 1999;1:358–367.
36. Nath R, Yue N, Liu L. On the depth of penetration of photons and electrons for intravascular brachytherapy. Cardiovasc Radiat Med 1999;1:72–79.
37. Li XA, Wang R, Yu C, Suntharalingam M. Beta versus gamma for catheter-based intravascular brachytherapy: Dosimetric perspectives in the presence of metallic stents and calcified plaques. Int J Radiat Oncol Biol Phys 2000;46:1043–1049.
38. Hanefeld C, et al. Dosimetric measurements in isolated human coronary arteries: Comparison of commercially available iridium(192) with strontium/yttrium(90) emitters. Circulation 2002;105:2493–2496.
39. Icru Report 26. International Commission on Radiation Units and Measurements, Bethesda (MD); 1977.
40. Fan P, et al. Effect of stent on radiation dosimetry in an in-stent restenosis model. Cardiovasc Radiat Med 2000;2:18–25.
41. Amols HI, Trichter F, Weinberger J. Intracoronary radiation for prevention of restenosis: Dose perturbations caused by stents. Circulation 1998;98:2024–2029.
42. Mourtada F, Horton JL. Dose perturbation of a novel cobalt chromium coronary stent on ^{32}P intravascular brachytherapy: A monte carlo study. Med Phys 2005;32:268–274.
43. Fox RA. Intravascular brachytherapy of the coronary arteries. Phys Med Biol 2002;47:R1–30.
44. Seltzer SM. Monte Carlo modeling for intravascular brachytherapy sources. In: Shope TB, editor. Intravascular brachytherapy and fluoroscopically guided interventions. Madison (WI): Medical Physics Publishing; 2002.
45. Jenkins TM, Nelson WR, Rindi A, editors. Monte carlo transport of electrons and photons. New York: Plenum Press; 1988.
46. Icru Report 56. International Commission On Radiation Units and Measurements, Bethesda (MA); 1997.
47. Berger MJ. Monte carlo calculations of the penetration and diffusion of fast charged particles. Alder B, Fernabach S, Rotenberg M, editors. Methods in computational physics. New York: Academic Press; 1963.
48. Niroomand-Rad A, et al. Radiochromic film dosimetry: Recommendations of aapm radiation therapy committee task group 55. Med Phys 1998;25:2093–2115.
49. Amin MN, et al. A comparison of polyacrylamide gels and radiochromic film for source measurements in intravascular brachytherapy. Br J Radiol 2003;76:824–831.
50. Collé R. Chemical digestion and radionuclide assay of tini-encapsulated 32p intravascular brachytherapy sources. Appl Rad Isotopes 1999;50:811–833.
51. Colle R, Zimmerman BE, Soares CG, Coursey BM. Determination of a calibration factor for the nondestructive assay of guidant 32P brachytherapy sources. Appl Radiat Isotopes 1999;50:835–841.
52. Collé R. On the radioanalytical methods used to assay stainless-steel-encapsulated, ceramic-based ^{90}Sr/^{90}Y intravascular brachytherapy sources. Appl Radiat Isotopes 2000;52:1–18.
53. Colle R. Activity characterization of pure-beta-emitting brachytherapy sources. Appl Radiat Isotopes 2002;56:331–336.
54. Waksman R, et al. Intravascular radiation therapy after balloon angioplasty of narrowed femoropopliteal arteries to prevent restenosis: Results of the paris feasibility clinical trial. J Vasc Interv Rad 2001;12:915–921.
55. Wang R, Li XA, Lobdell J. Monte carlo dose characterization of a new 90sr/90y source with balloon for intravascular brachytherapy. Med Phys 2003;30:27–33.
56. Bloch P, Bonan R, Wallner P, Lobdell J. Dosimetry for an ^{90}Sr/^{90}Y source train used for intravascular radiation of a hemodialysis graft. Cardiovasc Rad Med 2003;4:90–94.
57. Chan AW, et al. Carotid brachytherapy for in-stent restenosis. Catheter Cardiovasc Interv 2003;58:86–92.
58. Saito T, Waki K, Becker AE. Left atrial myocardial extension onto pulmonary veins in humans: Anatomic observations relevant for atrial arrhythmias. J Cardiovasc Electrophysiol 2000;11:888–894.
59. Degertekin M, et al. Sirolimus-eluting stent for treatment of complex in-stent restenosis: The first clinical experience. J Am Coll Cardiol 2003;41:184–189.
60. Morice MC, et al. A randomized comparison of a sirolimus-eluting stent with a standard stent for coronary revascularization. N Engl J Med 2002;346:1773–1780.
61. Abizaid A, et al. Sirolimus-eluting stents inhibit neointimal hyperplasia in diabetic patients. Insights from the ravel trial. Eur Heart J 2004;25:107–112.

62. Serruys PW, et al. Intravascular ultrasound findings in the multicenter, randomized, double-blind ravel (randomized study with the sirolimus-eluting velocity balloon-expandable stent in the treatment of patients with de novo native coronary artery lesions) trial. Circulation 2002;106:798–803.
63. Bailey SR. Drug-eluting stents have made brachytherapy obsolete. Curr Opin Cardiol 2004;19:598–600.
64. Kaluza GL, Raizner AE. Brachytherapy for restenosis after stenting for coronary artery disease: Its role in the drug-eluting stent era. Curr Opin Cardiol 2004;19:601–607.

Reading List

Waksman R, Serruys P. Handbook of Vascular Bracdhytherapy. 2nd ed. London: Martin Dunitz Ltd; 2000.
Waksman R. Vascular Brachytherapy. 2nd ed. Armonk (NY): Futura Publishing Company; 1999.
Hall EJ. Radiobiology for the Radiologist. 5th ed. Philadelphia: J.B. Lippincott; 2000.
Kutryk MJ, Serruys PW. Coronary Stenting Current Perspective. London: Martin Dunitz Ltd; 1999.
Balter S, Chan RC, Shope TB. Intravascular Brachytherapy, Fluoroscopically Guided Interventions. Madison: Medical Physics Publishing; 2002.
Leon MB, Mintz GS. Interventional Vascular Product Guide. London: Martin Dunitz Ltd; 1999.
Attix FH. Introduction to Radiological Physics and Radiation Dosimetry. New York: John Wiley & Sons, Inc.; 1986.

See also BRACHYTHERAPY, HIGH DOSAGE RATE; CORONARY ANGIOPLASTY AND GUIDEWIRE DIAGNOSTICS.

BRAIN ELECTRICAL ACTIVITY. See ELECTROENCEPHALOGRAPHY.

BURN WOUND COVERINGS. See SKIN SUBSTITUTE FOR BURNS, BIOACTIVE.

BYPASS, CORONARY. See VASCULAR GRAFT PROSTHESIS.

BYPASS, CARDIOPULMONARY. See HEART-LUNG MACHINES.